Electronic Journal of Statistics
Vol. 8 (2014) 96–129
ISSN: 1935-7524
DOI: 10.1214/14-EJS877

Data-based decision rules about the convexity of the support of a distribution

Pedro Delicado*

Dept. d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya
e-mail: pedro.delicado@upc.edu

Adolfo Hernández

Dept. de Estadística e Investigación Operativa II (Métodos de Decisión)
Universidad Complutense de Madrid
e-mail: a.hernandez@emp.ucm.es

and

Gábor Lugosi†

ICREA and Department of Economics
Pompeu Fabra University
e-mail: gabor.lugosi@upf.edu

Abstract: Given n independent, identically distributed random vectors in $\mathbb{R}^d$, drawn from a common density f, one wishes to find out whether the support of f is convex or not. In this paper we describe a decision rule which decides correctly for sufficiently large n, with probability 1, whenever f is bounded away from zero in its compact support. We also show that the assumption of boundedness is necessary. The rule is based on a statistic that is a second-order U-statistic with a random kernel. Moreover, we suggest a way of approximating the distribution of the statistic under the hypothesis of convexity of the support. The performance of the proposed method is illustrated on simulated data sets. As an example of its potential statistical implications, the decision rule is used to automatically choose the tuning parameter of ISOMAP, a nonlinear dimensionality reduction method.

MSC 2010 subject classifications: Primary 62G10; secondary 62H30.
Keywords and phrases: Discernibility between hypotheses, bootstrap subsampling, U-statistics, set estimation, ISOMAP, dimensionality reduction.

Received March 2013.

*Supported by Spanish Ministry of Education and Science and FEDER (MTM2010-14887).
†Supported by the Spanish Ministry of Science and Technology grant MTM2012-37195.

Contents

1. Introduction

Let X be a random vector with distribution μ on $\mathbb{R}^d$ having density f. The *support* of μ is defined as

$$S = \bigcap_{C \subset \mathbb{R}^d \text{ closed set}: \mu(C)=1} C. \tag{1}$$

(We also call S the support of the density f.) Let $X_1, \ldots, X_n$ be independent random vectors drawn from the distribution μ. In this paper we investigate the problem of testing whether the support S is a convex set or not. In other words, we consider the hypothesis testing problem in which the null and alternative hypotheses are

$$\begin{cases} H_0 : S \text{ is a convex set,} \\ H_1 : S \text{ is not a convex set.} \end{cases}$$

We are interested in finding tests—or, perhaps more adequately, *decision rules*—that decide correctly when the sample size is large. Formally, a decision rule is a sequence of functions $T_n : (\mathbb{R}^d)^n \to \{0, 1\}$. $T_n(X_1, \ldots, X_n) = 1$ is interpreted as a guess that f has a convex support while if $T_n(X_1, \ldots, X_n) = 0$, the decision rule suggests that the support is non-convex. A decision rule is *consistent* for a density f if it is correct eventually almost surely, that is, if

$$\mathbb{P}\left\{T_n(X_1, \ldots, X_n) \neq \mathbb{1}_{\{f \text{ has convex support}\}} \text{ for finitely many } n\right\} = 1.$$

Estimating the support (and other level sets) of a density from an i.i.d. sample has received considerable attention (see Baíllo, Cuevas and Justel (2000), Baíllo and Cuevas (2001), Cadre (2006), Biau, Cadre and Pelletier (2008), Cuevas and Fraiman (1997), Cuevas and Rodríguez-Casal (2004), Cuevas and Fraiman (2009), Cuevas (2009), Rodríguez-Casal (2007), Pateiro-López and Rodríguez-Casal (2009), Mason and Polonik (2009), Polonik (1995), Rigollet and Vert (2009), Scott and Nowak (2006), Steinwart, Hush and Scovel (2006), Tsybakov (1997), Vert and Vert (2006), Willett and Nowak (2007) for an incomplete but representative list of recent papers). However, as far as we know, no test has been proposed to decide whether the support of a density is convex or not. Apart

from its intrinsic interest, such a test has applications in the automatic choice of the tuning parameters of ISOMAP (isometric feature mapping; Tenenbaum, de Silva and Langford (2000)), a celebrated nonlinear dimensionality reduction method, as described in Section 5. In fact, it is this application that motivated our interest in the problem.

The main objective of this paper is to investigate the possibility of constructing consistent decision rules for the convexity of the support. We show that consistent decision rules (i.e., rules that decide correctly eventually almost surely) exist whenever f is bounded away from zero on its support and some other mild regularity conditions are satisfied. The rule, proposed in Section 2, is based on a statistic which is the average, over all pairs of points (X_i, X_j), of the distance of the closest data point to the mid-point $(X_i + X_j)/2$. We show that under the null hypothesis this average value converges to zero, in probability, while under the alternative, it stays bounded away from zero. This makes it possible to define a consistent decision rule. The difficulty of the analysis is that the proposed statistic is not a U-statistic since every summand depends not only on X_i and X_j but on all other data.

In Section 3 it is shown that it is impossible (in a well-defined sense described below) to design a decision rule that behaves asymptotically correctly for all bounded densities of bounded support. This shows that an assumption like the density being bounded away from zero on its support is necessary for consistent decision rules.

In Section 4, using the terminology of hypothesis testing, we describe some heuristics to approximate the distribution of the proposed statistic under the hypothesis of convexity of the support. Such approximations are essential in practice when the threshold for accepting or rejecting the null hypothesis needs to be adjusted for a given problem at a fixed sample size. We present numerical examples for illustration. Finally Section 5 illustrates by a numerical example how the decision rule is applied successfully in the automatic choice of the tuning parameter of ISOMAP.

2. A decision rule for the convexity of the support of a distribution

Let $X_1, \ldots, X_n$ be i.i.d. vectors drawn from the probability distribution μ on $\mathbb{R}^d$. We assume that μ is absolutely continuous with respect to the Lebesgue measure, with density f. Suppose that f has a support $S \subset \mathbb{R}^d$ and that there exists a constant $c > 0$ such that for every $x \in S$, $f(x) \geq c$. In this section we propose a test for the convexity of S. The main result of the section is that the decision rule is consistent, that is, regardless of whether S is convex or not, the rule decides correctly for sufficiently large sample sizes. For this we also need some mild regularity conditions detailed below.

The basic idea of the proposed test is the fact that a closed set $S \subseteq \mathbb{R}^d$ is convex if and only if for all $x, y \in S$, the mid-point $(x + y)/2$ is also in S. Thus, if the support S of f is convex, it is reasonable to expect that for each pair of observations (X_i, X_j), there is some other data point X_h close to the mid-point

$(X_i + X_j)/2$. On the other hand, if the support is not convex then we expect to have a large number of pairs (X_i, X_j) such that the closest point to $(X_i + X_j)/2$ is far away. Based on this intuition, we introduce the statistic

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \min_{h=1,\dots,n} \gamma(X_i, X_j, X_h) = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \gamma(X_i, X_j, X_{h_{(1)}(i,j)})$$

where $\gamma(X_i, X_j, X_h) = \|X_h - (X_i + X_j)/2\|^d$ and $h_{(1)}(i,j)$ denotes the index h for which $\|X_h - (X_i + X_j)/2\|$ is smallest. Observe that $\gamma(X_i, X_j, X_{h_{(1)}(i,j)}) \le \Delta_S^d$, where Δ_S denotes the diameter of S.

U_n resembles a U-statistic (see, e.g., Serfling (1980), Chapter 6) as it is a sum, over all pairs of points, of a function depending on the pair. However, U_n is not a U-statistic because the *kernel* $\gamma(X_i, X_j, X_{h_{(1)}(i,j)})$ depends not only on (X_i, X_j) but also on the rest of points X_h, which makes its analysis more complex. U-statistics with a random kernel were investigated by Schick (1997), but these results are not applicable to U_n as Schick deals with random kernels $\hat{k}_n(X_i, X_j; X_1, \dots, X_n)$ that converge (as $n \to \infty$) in some sense to a non-random kernel $k_n(X_i, X_j)$ for which the standard results on U-statistics apply. This is not the case for the kernel γ defining U_n.

In Propositions 1 and 2 below we show that, under a certain regularity condition, if the support is not convex, U_n stays bounded away from zero almost surely, while for convex S, its expectation converges to zero at a rate $(\log n/n)^{1/d}$ and it is concentrated around its mean. Thus, it makes sense to define the following rule:

$$\text{accept } H_0 \text{ if and only if } \quad U_n \le \tau_n$$

where $\tau_n \to 0$ but slower than $(\log n)/\sqrt{n}$. Indeed, this test is guaranteed to make the correct decision eventually, almost surely, whenever f is bounded from above and from below on its support. The regularity condition we require is the following.

Assumption 1. Assume that the topological boundary ∂S of S has zero Lebesgue measure.

Since the density f is supposed to be bounded away from zero on its support, the assumption is equivalent to saying that f is such that for almost every $x \in S$ there exists $\epsilon > 0$ such that $\operatorname{essinf}_{y:\|y-x\|<\epsilon} f(y) > 0$, see Lemma 4 in the appendix for the proof of this simple fact. Note that Assumption 1 is equivalent to the fact that S is Jordan measurable. If S is convex, the assumption is always satisfied, see Lang (1986).

The regularity assumption, together with the assumption that f is bounded away from zero on its support, exclude some pathological cases in which the statistical problem of deciding whether the support is convex is not only difficult, but also of questionable meaning. For example, Assumption 1 excludes cases such as a uniform density on a Cantor set of positive measure. As another illustration, consider the following example of a density over the real line. Let $r_1, r_2, \dots$ be an enumeration of all rational numbers. Then the set $A = \cup_{n \ge 1}(r_n -$

$2^{-(n+2)}, r_n + 2^{-(n+2)})$ has Lebesgue measure at most $1/2$ and we may define μ as the uniform distribution over A. Then the support S of μ is $\mathbb{R}$ (in particular, S is convex), yet the density vanishes everywhere except for a set of measure $1/2$. Our regularity assumptions exclude such pathological cases.

The next performance guarantee is the main result of this section:

Theorem 1. *Suppose that the support of f satisfies Assumption 1 and that there exist constants $0 < c < C$ such that $c \leq f(x) \leq C$ for all $x \in S$. Consider the test which accepts H_0 if and only if $U_n \leq \tau_n$ and suppose that τ_n is chosen such that*

$$\lim_{n\to\infty} \tau_n = 0 \quad \textit{and} \quad \lim_{n\to\infty} \frac{\tau_n n^{1/2}}{\log n} = \infty.$$

Then regardless of whether S is convex or not, with probability one, there exists an index N such that for all $n > N$ the test always decides correctly.

Remark 1. Of course, the density f is not uniquely defined as its value can be changed on a set of zero Lebesgue measure. The boundedness condition for f in the theorem should be interpreted such that f has a version that satisfies this condition. More precisely, we assume that $\operatorname{esssup}_{x\in S} f(x) \leq C$ and $\operatorname{essinf}_{x\in S} f(x) \geq c$. This comment applies throughout the whole paper.

The theorem is an immediate consequence of Propositions 1 and 2 below. As it is shown in Section 3, the condition of f being bounded away from zero cannot be dropped. However, we conjecture that the condition that f is bounded from above is not necessary.

Note that our choice of the function γ is far from being the only possibility that gives rise to a consistent decision rule. In particular, the d-th power of the norm may be replaced by any other positive power. However, the proposed choice has some advantages that we exploit in Section 4 in defining a bootstrap approximation of the distribution of U_n under the null hypothesis.

First we establish the asymptotic behavior of U_n under both the null and alternative hypotheses. We treat the simpler case, when S is not convex, first:

Proposition 1 (Asymptotic properties of U_n under H_1)**.** *Suppose that Assumption 1 is satisfied and that f is bounded away from zero on its support S. If S is not convex, then $\liminf_{n\to\infty} U_n > 0$ almost surely.*

Proof. For $z \in \mathbb{R}^d$ and $r > 0$, denote by $N(z, r)$ the open ball of radius r centered at z.

Suppose that S is not convex. Then there exist $x, y \in S$ such that $(x+y)/2 \notin S$. (The fact that we may take the mid-point of x and y follows from closedness of S.) Since S is closed, $(x + y)/2$ has a neighborhood entirely outside of S. Also, by Assumption 1, S equals the closure of the set $A = \{x : \exists \delta > 0 : \operatorname{essinf}_{y\in N(x,\delta)} f(y) > 0\}$ (see Lemma 5 in the appendix). This implies that there exist $x', y' \in A$ and $\epsilon > 0$ such that $N(x', \epsilon) \cup N(y', \epsilon) \subset A$ and $N((x' + y')/2, 2\epsilon) \cap S = \emptyset$. (Indeed, if $x \in A$ then we may take $x' = x$ otherwise any $x' \in A$ sufficiently close to x will do. y' is chosen similarly.) By assumption, there exists a constant $c > 0$ such that for every $x \in A$, $f(x) \geq c$. By the law of

large numbers, with probability one, there exists an index N such that for all $n > N$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in N(x',\epsilon)\}} \geq \frac{c}{2}\epsilon^d v_d \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in N(y',\epsilon)\}} \geq \frac{c}{2}\epsilon^d v_d,$$

where v_d is the volume of the d-dimensional unit Euclidean ball. On the other hand, clearly, if $X_i \in N(x',\epsilon)$ and $X_j \in N(y',\epsilon)$, $\gamma(X_i, X_j, h^{(1)}(i,j)) \geq \epsilon^d$. Since $N(x',\epsilon)$ and $N(y',\epsilon)$ are disjoint, if $n \geq N$, the number of such pairs (X_i, X_j) is at least $(nc\epsilon^d v_d)^2/4$ and therefore

$$\liminf_{n\to\infty} U_n \geq \frac{(c\epsilon^d v_d)^2 \epsilon^d}{2} > 0 \quad \text{almost surely.}$$

□

The next result shows that under the null hypothesis, the expected value of U_n goes to zero at a rate $(\log n/n)^{1/d}$ and it is very unlikely to exceed its expectation by more than $\log n/\sqrt{n}$. This result, combined with the Borel-Cantelli lemma, implies that for any sequence τ_n such that $\tau_n n^{1/d}/\log n \to \infty$, with probability one, $U_n < \tau_n$ for all sufficiently large n.

Proposition 2 (Asymptotic properties of U_n under H_0)**.** *Suppose that Assumption 1 is satisfied and that there exist constants $0 < c < C$ such that $c \leq f(x) \leq C$ for all $x \in S$. If S is convex, then there exists a constant K depending on c, C, and S such that, for all n and $q \geq 2$,*

$$\mathbb{E}U_n \leq K\left(\frac{\log n}{n}\right)^{1/d} \quad \textit{and} \quad \mathbb{E}\left[(U_n - \mathbb{E}U_n)_+^q\right] \leq \left(Kq^{3/2}\frac{\log n}{\sqrt{n}}\right)^q,$$

where $(\cdot)_+$ denotes positive part.

Proof. Note first that convexity of S implies that Assumption 1 holds and therefore the open set $A = \{x : \exists \delta > 0 : \operatorname{essinf}_{y\in N(x,\delta)} f(y) > 0\} \subset S$ is also convex. (To see this, consider $x, y \in A$ and $\lambda \in (0,1)$. Since A is open and S is convex, $\lambda x + (1-\lambda)y$ has a neighborhood entirely included in S. Since f is at least c at every point of S, this implies that $\lambda x + (1-\lambda)y \in A$.)

Since A is open, there exists an $x \in A$ and $\delta > 0$ such that $N(x,\delta)$ is contained in A.

Since f is assumed to be bounded away from zero on A which is convex, A must also be bounded. To see this, note that since A contains an open ball $N(x,\delta)$, if A was unbounded, for all $n > 0$ there would exist $x_n \in A$ with $\|x - x_n\| > n$. Since A is convex, it contains the convex hull of $N(x,\delta)$ and x_n whose volume is bounded from below by a positive constant (depending on δ and d) times n which contradicts the fact that A has a bounded Lebesgue measure.

Since S equals the closure of A (again by Lemma 5 in the appendix), S is compact.

By translating S if necessary, we may assume, without loss of generality, that $N(0,\delta) \subset A$. We may also assume, without loss of generality, that Δ_A, the diameter of A (and S), is equal to 1.

For all $\epsilon > 0$, define the ϵ-*interior* of A by

$$A_\epsilon = \{x \in A : B(x,\epsilon) \subset A\}$$

where $B(x,\epsilon)$ is the closed ball of radius ϵ centered at x. Note that A_ϵ is non-empty, open, and convex whenever $\epsilon \leq \delta$.

The reason we introduce A_ϵ is because to bound the expectation of U_n, we assume that both X_1 and X_2 are in the ϵ-interior of A and also show that the probability the assumption is not satisfied is small. More precisely, to bound the expected value, note first that for all $\epsilon \leq \delta$,

$$\begin{aligned}
\mathbb{E}U_n &= \mathbb{E}\left[\min_{h=1,\ldots,n} \gamma(X_1, X_2, X_h)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\min_{h=1,\ldots,n} \gamma(X_1, X_2, X_h) \mid X_1, X_2\right] \mathbb{1}_{\{X_1 \text{ or } X_2 \in A \setminus A_\epsilon\}}\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\min_{h=1,\ldots,n} \gamma(X_1, X_2, X_h) \mid X_1, X_2\right] \mathbb{1}_{\{X_1, X_2 \in A_\epsilon\}}\right]. \qquad (2)
\end{aligned}$$

Since $\gamma(X_1, X_2, X_h) \leq 1$, the first term on the right-hand side may be bounded by

$$\mathbb{P}\{X_1 \text{ or } X_2 \in A \setminus A_\epsilon\} \leq 2\mathbb{P}\{X_1 \in A \setminus A_\epsilon\} \leq 2C \,\mathrm{Vol}\,(A \setminus A_\epsilon)$$

where Vol denotes the d-dimensional Lebesgue measure. To bound the volume of the boundary region, first observe that since S is the closure of A, $\mathrm{Vol}(A \setminus A_\epsilon) = \mathrm{Vol}(S \setminus A_\epsilon)$. Next we show that there exists a constant $\kappa_S > 0$, depending on S, such that for all $\epsilon < \delta/4$, $S \subset A_\epsilon \oplus N(0, \kappa_S \epsilon)$ where $\oplus$ denotes Minkowski sum. This may be seen as follows: since $N(0,\delta) \subset A$, every $x \in S \setminus A_\epsilon$ is not in $N(0,\delta/2)$. Let

$$\theta(x) = \inf\{\theta > 0 : N(0,\delta/2) \subset x + C(-x,\theta)\}$$

denote the infimum of the angle of any cone centered at x that includes $N(0,\delta/2)$ where for $x \in \mathbb{R}^d$, a cone of angle θ is defined as

$$C(x,\theta) = \left\{y \in \mathbb{R}^d : \frac{x^T y}{\|x\|\|y\|} \geq \cos(\theta)\right\}$$

(see Figure 1). Since S is compact and $\theta(x)$ is positive and continuous, $\theta = \inf_{x \in S \setminus A_\epsilon} \theta(x) > 0$. Let $x \in A \setminus A_\epsilon$ and define $y = ax$ where $a = \sup\{\alpha \in (0,1) : N(\alpha x, \epsilon) \subset x + C(-x, \theta(x))\}$. In words, y is the point on the segment joining x and 0 such that $N(y,\epsilon)$ "just fits" in the cone $x + C(-x,\theta(x))$, see Figure 2. (Such a point exists by the definition of $\theta(x)$ and since $\epsilon < \delta/4$.) Note that $N(y,\epsilon)$ lies in the convex hull of $\{x\} \cup N(0,\delta)$ and therefore, by convexity of A, $N(y,\epsilon) \subset A$. Since A is open, this implies that $B(y,\epsilon) \subset A$ and therefore $y \in A_\epsilon$.

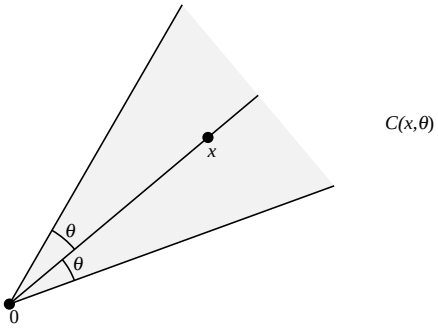

FIG 1. *A cone of angle θ.*

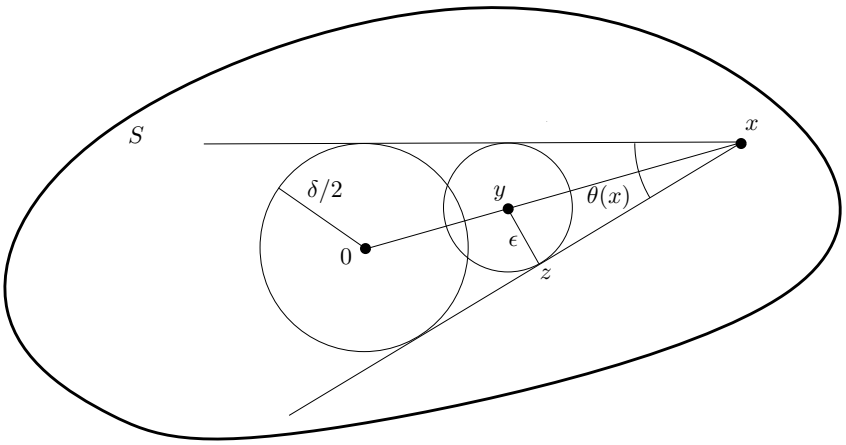

FIG 2. *The point y in the proof of Proposition 2.*

Consider now any straight line containing x and tangent to $N(y,\epsilon)$ at, say, point z. The right-angle triangle formed by x, y, and z is such that its hypotenuse is the segment $[x,y]$, and its leg $[y,z]$ has length ϵ. Since the angle of the triangle at vertex x equals $\theta(x)$, we have that $\|x-y\| \leq \epsilon/\sin(\theta(x)) \leq \epsilon/\sin(\theta)$. Therefore, we may take $\kappa_S = 1/\sin(\theta)$. Therefore, for all $\epsilon < \delta/\kappa_S$,

$$\begin{aligned}
\mathrm{Vol}(S) &= \mathrm{Vol}(A) \leq \mathrm{Vol}(A_\epsilon \oplus N(0,\kappa_S\epsilon)) \\
&\leq \mathrm{Vol}\left(A_\epsilon \oplus \frac{\kappa_S\epsilon}{\delta-\kappa_S\epsilon}A_\epsilon\right) \quad (\text{since } N(0,\delta-\kappa_S\epsilon) \subset A_\epsilon) \\
&= \mathrm{Vol}\left(\left(1+\frac{\kappa_S\epsilon}{\delta-\kappa_S\epsilon}\right)A_\epsilon\right) \quad (\text{since } A_\epsilon \text{ is convex}) \\
&= \mathrm{Vol}(A_\epsilon)\left(1+\frac{\kappa_S\epsilon}{\delta-\kappa_S\epsilon}\right)^d \\
&\leq \mathrm{Vol}(A_\epsilon)\left(1+\frac{4d\kappa_S}{\delta}\epsilon\right) \quad \text{if} \quad \epsilon \leq \min\left(\frac{\delta}{2\kappa_S}, \frac{2\delta}{d2^{d-2}}\right),
\end{aligned}$$

where the last inequality follows from Taylor's theorem. Thus,

$$\mathrm{Vol}\,(A \setminus A_\epsilon) = \mathrm{Vol}(A) - \mathrm{Vol}(A_\epsilon) \leq \mathrm{Vol}(A_\epsilon)\frac{4d\kappa_S}{\delta}\epsilon \leq \mathrm{Vol}(S)\frac{4d\kappa_S}{\delta}\epsilon$$

Hence, we have shown that the first term on the right-hand side of (2) may be bounded as

$$\mathbb{E}\left[\mathbb{E}\left[\min_{h=1,\ldots,n}\gamma(X_1,X_2,X_h) \mid X_1,X_2\right]\mathbb{1}_{\{X_1 \text{ or } X_2 \in A\setminus A_\epsilon\}}\right] \leq K_1\epsilon$$

where $K_1 = 2C\,\mathrm{Vol}(S)\frac{4d\kappa_S}{\delta}$ is a constant depending on S and C only.

It remains to bound the second term on the right-hand side of (2). To this end, suppose $\epsilon \leq \delta$. In the event that $X_1, X_2 \in A_\epsilon$,

$$\begin{aligned}
&\mathbb{E}\left[\min_{h=1,\ldots,n}\gamma(X_1,X_2,X_h) \mid X_1,X_2\right] \\
&= \int_0^1 \mathbb{P}\left\{\min_{h=1,\ldots,n}\gamma(X_1,X_2,X_h) > t \mid X_1,X_2\right\} dt \\
&= \int_0^1 \mathbb{P}\left\{\gamma(X_1,X_2,X_3) > t \mid X_1,X_2\right\}^{n-2} dt \\
&\leq \epsilon + \left(1 - cv_d\epsilon^d\right)^{n-2} \\
&\quad \text{(since } \mathbb{P}\left\{\gamma(X_1,X_2,X_3) \leq t \mid X_1,X_2\right\} \geq cv_d\epsilon^d \text{ when } t \geq \epsilon \text{ and } X_1,X_2 \in A_\epsilon\text{)} \\
&\leq \epsilon + e^{-cv_d\epsilon(n-2)}.
\end{aligned}$$

Summarizing, we have proved that, for all $\epsilon \leq \min(\frac{\delta}{2}, \frac{2\delta}{d2^{d-2}})$,

$$\mathbb{E}U_n \leq (K_1+1)\epsilon + e^{-cv_d\epsilon^d(n-2)}.$$

Choosing $\epsilon = (\log n/((n-2)cv_d))^{1/d}$ completes the proof of the bound for the expected value of U_n.

To bound the higher moments of U_n, we apply a general moment inequality for functions of independent random variables of Boucheron et al. (2005, Theorem 3) which states that, for every $q \geq 2$,

$$\mathbb{E}\left[(U_n - \mathbb{E}U_n)_+^q\right] \leq (\kappa q)^{q/2}\mathbb{E}\left[\left(\sum_{k=1}^n (U_n - U_{n,k})^2\right)^{q/2}\right]$$

where $\kappa = \sqrt{e}/(2\sqrt{e}-2) < 1.271$ and

$$U_{n,k} = \frac{1}{\binom{n}{2}}\sum_{(i,j):i<j,i\neq k,j\neq k}\gamma(X_i,X_j,X_{h^k_{(1)}(i,j)})$$

with $h^k_{(1)}(i,j) \in \{1,\ldots,n\}\setminus\{k\}$ defined as $h_{(1)}(i,j)$ but with X_k omitted from the sample.

Thus, we need to study the effect of removing the point X_k from the sample on the value of U_n. Clearly,

$$U_n - U_{n,k} = \frac{1}{\binom{n}{2}} \sum_{j:j\neq k} \gamma(X_k, X_j, X_{h_{(1)}(k,j)}) + \frac{1}{\binom{n}{2}} \sum_{(i,j):i<j,i\neq k,j\neq k} \left(\gamma(X_i, X_j, X_{h_{(1)}(i,j)}) - \gamma(X_i, X_j, X_{h^k_{(1)}(i,j)})\right)$$

The first term on the right-hand side is non-negative and bounded by $2/n$. At the same time, every term in the second sum on the right-hand is in $[-1, 0]$ and is not zero only if $h_{(1)}(i, j) = k$. This implies that

$$(U_n - U_{n,k})^2 \leq \max\left(\frac{4}{n^2}, \left(\frac{1}{\binom{n}{2}} \sum_{(i,j):i<j,i\neq k,j\neq k} \mathbb{1}_{\{h_{(1)}(i,j)=k\}}\right)^2\right).$$

Thus, denoting by $N_k = \sum_{(i,j):i<j,i\neq k,j\neq k} \mathbb{1}_{\{h_{(1)}(i,j)=k\}}$ the number of pairs (X_i, X_j) of points in the sample for which X_k is the closest point to $(X_i+X_j)/2$, we have

$$\begin{aligned}\mathbb{E}\left[(U_n - \mathbb{E}U_n)_+^q\right] &\leq (\kappa q)^{q/2}\left(\left(\frac{4}{n}\right)^{q/2} + \frac{1}{\binom{n}{2}^q}\mathbb{E}\left[\left(\sum_{k=1}^n N_k^2\right)^{q/2}\right]\right)\\ &\leq (\kappa q)^{q/2}\left(\left(\frac{4}{n}\right)^{q/2} + \frac{n^{q/2}}{\binom{n}{2}^q}\mathbb{E}\left[\max_{k=1,\ldots,n} N_k^q\right]\right).\end{aligned}$$

Thus, it suffices to find suitable upper bounds for the moments of $\max_k N_k$. To this end, note that since $N_k \leq \binom{n-1}{2}$, for all $t > 0$,

$$\begin{aligned}\mathbb{E}\left[\max_{k=1,\ldots,n} N_k^q\right] &\leq \binom{n-1}{2}^q \mathbb{P}\left\{\max_{k=1,\ldots,n} N_k^q > (nt)^q\right\} + (nt)^q\\ &\leq n^{2q+1}\mathbb{P}\left\{N_1 > nt\right\} + (nt)^q.\end{aligned}$$

Next we write

$$N_1 = \frac{1}{2}\sum_{i=2}^n N_{1,i}$$

where $N_{1,i} = \sum_{j:j\neq i}^n \mathbb{1}_{\{h_{(1)}(i,j)=1\}}$ is the number of points $j \neq i$ such that $h_{(1)}(i, j) = 1$. Then

$$\mathbb{P}\left\{N_1 > nt\right\} \leq n \max_{i>1} \mathbb{P}\left\{N_{1,i} > t\right\}$$

It remains to bound $\mathbb{P}\{N_{1,i} > t\}$. In Lemma 1 below we show that there exists a constant K depending on c and the set S such that, for all n and $t > 0$,

$$\mathbb{P}\{N_{1,i} > t\} \leq Ke^{-Kn\epsilon_n^d}.$$

Putting everything together, we obtain that there exists a constant K (possibly different from the one above) such that

$$\mathbb{E}\left[(U_n - \mathbb{E}U_n)_+^q\right] \leq K^q q^{q/2}\left(n^{-q/2} + n^{q/2+2}e^{-Kn\epsilon_n^d} + n^{q/2}\epsilon_n^{dq}\right).$$

Choosing $\epsilon_n = K'((q\log n)/n)^{1/d}$ for a sufficiently large K', the upper bound becomes

$$\mathbb{E}\left[(U_n - \mathbb{E}U_n)_+^q\right] \leq K^q q^{3q/2}\frac{\log^q n}{n^{q/2}}$$

as desired. □

In the proof above we have used the following auxiliary result:

Lemma 1. *For $i = 2, 3, \ldots, n$, define $N_{1,i} = \sum_{j:j\neq i}^{n} \mathbb{1}_{\{h_{(1)}(i,j)=1\}}$. Then there exists a constant K depending on c and the set S such that, for all n and $t > 0$,*

$$\mathbb{P}\{N_{1,i} > t\} \leq Ke^{-Kn\epsilon_n^d}.$$

Proof. In the proof we condition on the value of $X_1 = x_1$ and consider two different cases: the first, somewhat simpler, case is when x_1 falls in the ϵ_n-interior of A (with ϵ_n defined below). The case when X_1 is closer than ϵ_n to the boundary can be handled by a similar argument, though one should proceed with some care.

Let $\epsilon_n = ((K_2 q\log n)/n)^{1/d}$ for some constant K_2 specified below. Recall that A_{ϵ_n} denotes the ϵ_n-interior of A.

Case 1. $x_1 \in A_{\epsilon_n}$.

By Lemma 5.5 in Devroye, Györfi and Lugosi (1996), $\mathbb{R}^d$ can be covered by $\rho_d = \lceil(1 + 2/\sqrt{2-\sqrt{3}})^d\rceil$ cones of angle $\pi/6$, that is, there exist ρ_d points $z_1, \ldots, z_{\rho_d}$ such that $\cup_{i=1}^{\rho_d} C(z_i, \pi/6) = \mathbb{R}^d$.

Now consider ρ_d cones $C_1, \ldots, C_{\rho_d}$ of angle $\pi/6$ centered at x_1 that cover $\mathbb{R}^d$. Consider the data points falling in each cone and mark the nearest neighbor of x_1. Let $X_{NN}^{(1)}, \ldots, X_{NN}^{(\rho_d)}$ denote these nearest neighbors. Let $R = \max_{j=1,\ldots,\rho_d} \|x_1 - X_{NN}^{(j)}\|$ be the distance of x_1 and the farthest of these nearest neighbors and define, for each $j = 1, \ldots, \rho_d$,

$$Z_{NN}^{(j)} = x_1 + R\frac{X_{NN}^{(j)} - x_1}{\left\|X_{NN}^{(j)} - x_1\right\|}$$

as the projection of $X_{NN}^{(j)}$ to the surface of the ball centered at x_1, of radius R.

Now for each $j = 1, \ldots, \rho_d$, let

$$H_j = \left\{x \in \mathbb{R}^d : \|x - x_1\| \leq \|x - Z_{NN}^{(j)}\|\right\}$$

be the half-space containing x_1 defined by the bisecting hyperplane between x_1 and $Z_{NN}^{(j)}$. The intersection $P = \cap_{j=1}^{\rho_d} H_j$ defines a convex polytope with

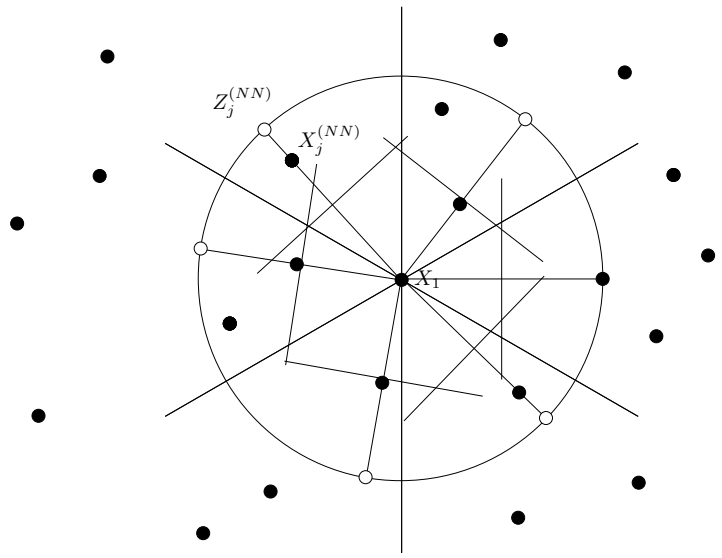

FIG 3. *Nearest neighbors $X_{NN}^{(j)}$ in each cone, their projection $Z_{NN}^{(j)}$ to the sphere of radius R, and the convex polytope P defined by the bisecting hyperplanes.*

ρ_d facets, see Figure 3. The key observation is that if $h_{(1)}(i,j) = 1$ for a pair (X_i, X_j), then $(X_i + X_j)/2 \in P$, otherwise one of the $X_{NN}^{(j)}$ would be closer to $(X_i + X_j)/2$ than x_1.

It is easy to see (and this is the second key observation) that $P \subset B(x_1, R)$ where $B(x_1, R)$ is the closed ball of radius R centered at x_1. Thus, for every $j \neq i$ such that $h_{(1)}(i,j) = 1$, we have $(X_i + X_j)/2 \in B(x_1, R)$ which is equivalent to saying that

$$X_j \in B(X_i + 2(x_1 - X_i), 2R).$$

Thus,

$$N_{1,i} \leq \sum_{j \geq 2 : j \neq i} \mathbb{1}_{\{X_j \in B(X_i + 2(x_1 - X_i), 2R)\}}. \tag{3}$$

We use the decomposition

$$\mathbb{P}\{N_{1,i} > t \mid X_1 = x_1\} \leq \mathbb{P}\{N_{1,i} > t \mid R < \epsilon_n, X_1 = x_1\} + \mathbb{P}\{R \geq \epsilon_n \mid X_1 = x_1\}$$

Note that given $R < \epsilon_n$, by (3), $N_{1,i}$ is dominated by a binomial random variable with parameters $n-1$ and $\mu(B(X_i + 2(x_1 - X_i), 2\epsilon_n)) \geq c v_d (2\epsilon_n)^d$ where v_d is the volume of the d-dimensional unit Euclidean ball. Therefore, choosing $t = 2(n-1)cv_d(2\epsilon_n)^d$, and setting $c_1 = \log(4/e)cv_d 2^{d+1}$, by a standard estimate for the tail of the binomial distribution,

$$\mathbb{P}\left\{N_{1,i} > 2(n-1)cv_d(2\epsilon_n)^d \mid R < \epsilon_n, X_1 = x_1\right\} \leq e^{-c_1(n-1)\epsilon_n^d}.$$

It remains to bound $\mathbb{P}\{R \geq \epsilon_n \mid X_1 = x_1\}$. Clearly,

$$\begin{aligned}\mathbb{P}\{R \geq \epsilon_n \mid X_1 = x_1\} &= \mathbb{P}\left\{\max_{j=1,\ldots,\rho_d} \|x_1 - X_{NN}^{(j)}\| \geq \epsilon_n \mid X_1 = x_1\right\} \\ &\leq \sum_{j=1}^{\rho_d} \mathbb{P}\left\{\|x_1 - X_{NN}^{(j)}\| \geq \epsilon_n \mid X_1 = x_1\right\}.\end{aligned}$$

Since $x_1 \in A_{\epsilon_n}$, we have $\inf_{x \notin A} \|x_1 - x\| > \epsilon_n$, and therefore

$$\mathbb{P}\left\{\|x_1 - X_{NN}^{(j)}\| \geq \epsilon_n \mid X_1 = x_1\right\} = \mathbb{E}\left[(1 - \mu(B(x_1, \epsilon_n) \cap C_j))^{n-1}\right]$$

Since x_1 is at least ϵ_n away from the complement of A and the density f is bounded from below by c on A,

$$\mu(B(x_1, \epsilon_n) \cap C_j) \geq \frac{c v_d}{\rho_d} \epsilon_n^d$$

Therefore,

$$\mathbb{P}\left\{\|x_1 - X_{NN}^{(j)}\| \geq \epsilon_n \mid X_1 = x_1\right\} \leq \left(1 - c_2 \epsilon_n^d\right)^{n-1} \leq e^{-c_2(n-1)\epsilon_n^d}$$

where $c_2 = c v_d / \rho_d$.

Putting everything together, we have

$$\mathbb{P}\{R \geq \epsilon_n \mid X_1 = x_1\} \leq \rho_d e^{-c_2(n-1)\epsilon_n^d}$$

and therefore

$$\mathbb{P}\{N_{1,i} > t \mid X_1 = x_1\} \leq e^{-c_1(n-1)\epsilon_n^d} + \rho_d e^{-c_2(n-1)\epsilon_n^d}.$$

Case 2. $x_1 \notin A_{\epsilon_n}$.

It remains to handle the case when x_1 is not in the ϵ_n-interior of A. Suppose that n is so large that $\epsilon_n < \delta/2$. The key observation is that there exists $\alpha \in (0, \pi/12]$ such that for all $x_1 \in A$, there exists a cone C_1 centered at x_1, of angle α such that $C_1 \cap B(x_1, \epsilon_n) \subset A$. (This follows by a similar argument that we have used earlier: first note that every $x_1 \in A \setminus A_{\epsilon_n}$ is not in $B(0, \delta/2)$—recall that $B(0, \delta) \subset A$. By convexity, the smallest cone centered at x_1 that includes $B(0, \delta/2)$ satisfies the required property. The smallest angle of all such cones over $x_1 \in A \setminus A_{\epsilon_n}$ is bounded away from zero by compactness of S.)

Now fix $x_1 \in A \cap A_{\epsilon_n}^c$. Cover $\mathbb{R}^d$ by a minimal number of cones $C_1, \ldots, C_{N_\alpha}$ centered at x_1 of angle α such that one of the cones C_1 is such that $C_1 \cap B(x_1, \epsilon_n) \subset A$. (Note that $N_\alpha \leq (1 + 1/(\sin(\alpha/2)))^d - 1$.) Observe that if each cone C_i with $C_i \cap B(x_1, \epsilon_n) \subset A$ contains at least one data point then for every pair (X_i, X_j) such that $h_{(1)}(i, j) = 1$, $(X_i + X_j)/2 \in B(x_1, \epsilon_n) \cap A$ (see Figure 4). Then the same argument as in the case of $x_1 \in A_{\epsilon_n}$ carries over with the only

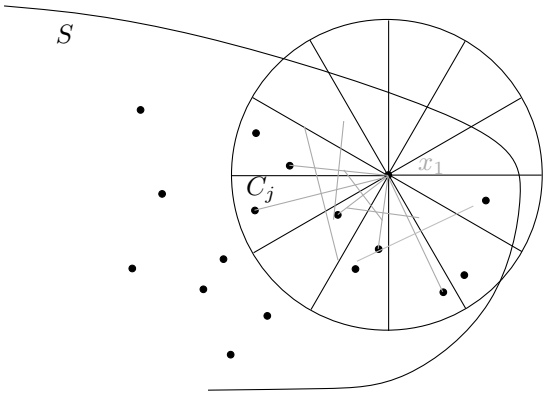

FIG 4. *When $C_j \cap B(x_1, \epsilon_n)$ is contained in S then it is very likely to contain a data point. If the angle of the cones is less than $\pi/12$, the convex set defined by the intersection of S with the bisecting half-spaces is contained in $B(x_1, \epsilon_n)$.*

difference that instead of cones of angle $\pi/6$ now we have cones of angle α and we obtain that there exists a constant K, depending on c and α such that

$$\mathbb{P}\{N_{1,i} > t \mid X_1 = x_1\} \leq Ke^{-Kn\epsilon_n^d}.$$

Since the above estimate holds independently of what x_1 is, we have established that

$$\mathbb{P}\{N_{1,i} > t\} \leq Ke^{-Kn\epsilon_n^d}.$$

□

3. On the non-discernibility of support convexity

The purpose of this section is to show that without the assumption that the density is bounded away from zero on its support, Theorem 1 is not true. Without further assumptions, it is impossible to decide whether the support of a density is convex. In order to formalize this statement, we recall the notion of *discernibility* introduced by Dembo and Peres (1994) (see also Devroye and Lugosi (2002)).

Let $\mathcal{F}$ and $\mathcal{G}$ be two disjoint sets of densities on $\mathbb{R}^d$. Let $X_1, \ldots, X_n$ be independent random vectors drawn according to a density $f \in \mathcal{F} \cup \mathcal{G}$. Based on these data, one tries to decide whether $f \in \mathcal{F}$ or not. Recall from Section 1 that a decision rule is a sequence of functions $T_n : (\mathbb{R}^d)^n \to \{0, 1\}$. $T_n(X_1, \ldots, X_n) = 1$ means that the rule guesses that $f \in \mathcal{F}$ while if $T_n(X_1, \ldots, X_n) = 0$, the decision rule thinks that $f \in \mathcal{G}$. A decision rule is consistent if, for every $f \in \mathcal{F} \cup \mathcal{G}$ it is correct eventually almost surely, that is, if

$$\mathbb{P}\left\{T_n(X_1, \ldots, X_n) \neq \mathbb{1}_{\{f \in \mathcal{F}\}} \text{ for finitely many } n\right\} = 1.$$

We say that the pair $(\mathcal{F}, \mathcal{G})$ is *discernible* if there exists a consistent decision rule.

Theorem 1 shows that if $\mathcal{F}$ is the class of densities that are bounded, bounded away from zero, and have convex support and $\mathcal{G}$ is the class of bounded densities, bounded away from zero with non-convex support (in both cases satisfying Assumption 1), then the pair $(\mathcal{F}, \mathcal{G})$ is discernible.

The main result of this section implies that the sets of all uniformly bounded densities with convex and non-convex support are not discernible. In other words, every decision rule must fail infinitely often for some density. Thus, the assumption of boundedness (from below) of the densities is necessary in Theorem 1 or at least needs to be replaced by another assumption. This is true even if we only consider densities on $\mathbb{R}$ with support in $[0, 1]$:

Theorem 2. *Let $\mathcal{F}$ be the class of all densities on $\mathbb{R}$ bounded by 2 with support $[0, 1]$, and let $\mathcal{G}$ be the class of all densities on $\mathbb{R}$ bounded by 2, satisfying Assumption 1, whose support is a non-convex subset of $[0, 1]$. Then the pair $(\mathcal{F}, \mathcal{G})$ is not discernible.*

A general impossibility theorem that gives sufficient conditions for a pair $(\mathcal{F}, \mathcal{G})$ to be non-discernible was given by Devroye and Lugosi (2002, Theorem 8). However, their theorem does not seem to apply here and we need a separate proof. We crucially use the following basic and well-known fact (see, e.g., Devroye and Györfi (2002)): if X and Y are real random variables with density f and g, respectively, then there exists a coupling (i.e., a joint distribution of (X, Y) with marginal densities f, g) such that $\mathbb{P}\{X \neq Y\} = (1/2) \int |f - g|$.

Proof of Theorem 2. To prove the theorem, we assume that the pair $(\mathcal{F}, \mathcal{G})$ is discernible, that is, there exists a consistent decision rule T_n. We construct subclasses $\mathcal{A} \subset \mathcal{F}$ and $\mathcal{B} \subset \mathcal{G}$ such that for any consistent decision rule T_n, there is a density ϕ in the L_1-closure of $\mathcal{A} \cup \mathcal{B}$ such that if $X_1, X_2, \ldots$ are distributed as ϕ then, with probability at least $1/2$, T_n changes decision infinitely many times, thus arriving at a contradiction.

Consider the following subclasses: let $\mathcal{B} = \{g_k : k = 1, 2, \ldots\}$ be the set of densities on $[0, 1]$ with non-convex support defined by

$$g_k(x) = \begin{cases} 2 & \text{if } x \in \left(\sum_{i=1}^{k} 2^{-i}, \sum_{i=1}^{k} 2^{-i} + 2^{-(k+2)}\right) \\ 0 & \text{if } x \in \left(\sum_{i=1}^{k} 2^{-i} + 2^{-(k+2)}, \sum_{i=1}^{k} 2^{-i} + 2^{-(k+1)}\right) \\ 1 & \text{otherwise.} \end{cases}$$

We also define the set $\mathcal{A} = \{f_{j,k} : j, k = 1, 2, \ldots\}$ of densities with full support $[0, 1]$ by

$$f_{k,j}(x) = \begin{cases} 2 - 2^{-j} & \text{if } x \in \left(\sum_{i=1}^{k} 2^{-i}, \sum_{i=1}^{k} 2^{-i} + 2^{-(k+2)}\right) \\ 2^{-j} & \text{if } x \in \left(\sum_{i=1}^{k} 2^{-i} + 2^{-(k+2)}, \sum_{i=1}^{k} 2^{-i} + 2^{-(k+1)}\right) \\ 1 & \text{otherwise.} \end{cases}$$

Assume that there exists a consistent decision rule T_n. Then for any density $f \in \mathcal{A}$, and almost all ω, there exists an integer $N(\omega)$ such that

$$T_n(X_1, \ldots, X_n) = 1 \quad \text{if } n > N(\omega)$$

and for any density $f \in \mathcal{B}$, and almost all ω there exists an integer $N(\omega)$ such that

$$T_n(X_1, \ldots, X_n) = 0 \quad \text{if } n > N(\omega).$$

Let $\delta_k = 2^{-k-2}$, $k = 1, 2, \ldots$. Let $\phi_1 \in \mathcal{A}$ be arbitrary. For concreteness, we may take $\phi_1 = f_{1,1}$. Let $X_1^{(1)}, X_2^{(1)}, \ldots$ be independent random variables with density ϕ_1. Since T_n is consistent, there exists an integer N_1 such that

$$\mathbb{P}\left\{T_n(X_1^{(1)}, \ldots, X_n^{(1)}) = 1 \text{ for all } n \geq N_1\right\} > 1 - \delta_1$$

(see, e.g., Royden (1968, p. 70, Problem 23.a)). Choose $\epsilon_1 > 0$ such that $(1 - \epsilon_1)^{N_1} > 1 - \delta_1$, and choose a density $\phi_2 \in \mathcal{B}$ with non-convex support such that $\phi_2(x) = \phi_1(x)$ for all $x \leq 1 - \epsilon_1$. For example, $\phi_2 = g_{k_2}$ with $k_2 \geq \log_2(1/\epsilon_1)$ will do. Then $\int |\phi_1 - \phi_2| \leq 2\epsilon_1$.

Now let $X_1^{(2)}, X_2^{(2)}, \ldots$ be independent random variables with density ϕ_2. Since $\phi_2 \in \mathcal{B}$, there exists an integer $N_2 > N_1$ such that

$$\mathbb{P}\left\{T_n(X_1^{(2)}, \ldots, X_n^{(2)}) = 0 \text{ for all } n \geq N_2\right\} > 1 - \delta_2.$$

Next we choose $\epsilon_2 > 0$ such that $(1 - \epsilon_2)^{N_2} > 1 - \delta_2$, and consider a convex-support density $\phi_3 \in \mathcal{A}$ which agrees with $\phi_2 = g_{k_2}$ for all $x \leq \sum_{i=1}^{k_2} 2^{-i}$ and $\int |\phi_2 - \phi_3| \leq 2\epsilon_2$. We may take $\phi_3 = f_{k_2, j_3}$ for any j_3 such that $j_3 \geq \log_2(2^{-(k_2+2)}/\epsilon_2)$.

We may continue this procedure such that $\phi_m \in \mathcal{A}$ for all odd m and $\phi_m \in \mathcal{B}$ for all even m, and $N_1 < N_2 < \cdots$ are chosen such that

$$\mathbb{P}\left\{T_n(X_1^{(m)}, \ldots, X_n^{(m)}) = \mathbb{1}_{\{m \text{ is odd}\}} \text{ for all } n \geq N_m\right\} > 1 - \delta_m.$$

The sequence of densities ϕ_m converges in L_1 to a density ϕ such that $\int |\phi_m - \phi| < 2\epsilon_m$ and ϕ agrees with ϕ_m for all $x \leq \sum_{i=1}^{k_{m-1}} 2^{-i}$ (which converges to 1 as $m \to \infty$).

Now let $X_1, X_2, \ldots$ be independent random variables drawn according to the density ϕ. Then, according to the coupling mentioned before the proof, there exist random variables

$$X_1^{(1)}, \ldots, X_{N_1}^{(1)}, X_1^{(2)}, \ldots, X_{N_2}^{(2)}, X_1^{(3)}, \ldots, X_{N_3}^{(3)}, \ldots$$

such that $X_i^{(m)}$ is distributed according to ϕ_m,

$$X_1^{(1)}, \ldots, X_{N_1}^{(1)}, X_{N_1+1}^{(2)}, \ldots, X_{N_2}^{(2)}, X_{N_2+1}^{(3)}, \ldots, X_{N_3}^{(3)}, \ldots$$

are independent, and $\mathbb{P}\{X_i \neq X_i^{(m)}\} < \epsilon_m$ for all $i \leq N_m$.

Then

$$\mathbb{P}\left\{X_1 = X_1^{(1)}, \ldots, X_{N_1} = X_{N_1}^{(1)}\right\} \geq (1 - \epsilon_1)^{N_1} > 1 - \delta_1,$$

and therefore

$$\mathbb{P}\left\{T_{N_1}(X_1,\ldots,X_{N_1})=1\right\}>1-2\delta_1.$$

Similarly, for each m,

$$\mathbb{P}\left\{X_1=X_1^{(m)},\ldots,X_{N_m}=X_{N_m}^{(m)}\right\}\geq(1-\epsilon_1)^{N_{m-1}}>1-\delta_{m-1},$$

and therefore

$$\mathbb{P}\left\{T_{N_m}(X_1,\ldots,X_{N_m}))=\mathbb{1}_{\{m \text{ is odd}\}} \text{ for all } m=1,2,\ldots\right\}>1-\sum_{m=1}^{\infty}2\delta_m=\frac{1}{2}.$$

Hence, with probability 1/2, the decision rule changes its decision infinitely often, concluding the proof. □

4. Data-based heuristics for calibrating the decision rule

Theorem 1 shows consistency of the rule that decides that the support is convex if and only if $U_n \leq \tau_n$ and it provides an asymptotic criterion to select τ_n. Nevertheless this result is not directly applicable in practice: if the sequence τ_n verifies the assumptions in Theorem 1 then so does $\tau_n^* = k\tau_n$ for any positive k, but it can be the case that $U_n \leq \tau_n$ whereas $U_n > \tau_n^*$ for a fixed n. In order to address this question, we find it convenient to use the standard terminology of hypothesis testing where the null hypothesis is that the underlying distribution has convex support.

An objective way of selecting τ_n is needed so that it is possible to control the probability of either of the two possible errors: Deciding that the support is not a convex set when indeed it is (type I error), and deciding that the support is convex when it is not the case (type II error).

The main difficulty is that the optimal value of the threshold τ_n depends on the unknown set S and the distribution μ. Therefore a mechanism is required to obtain a value for $\tau_n = \tau_n(S,\mu)$ that should be valid for any S and μ in a large class of distributions. The situation is similar to that appearing in the usual practice in bootstrap methods, where the distribution of a given statistic T is unknown and moreover it depends on the specific distribution of the data under study. In this context bootstrap methods are used to approach the specific distribution of T for every particular case.

We present a bootstrap-type approximation of the distribution of U_n under the hypothesis of support convexity (this procedure is also referred as calibration of the null distribution of U_n). We provide a heuristic justification that this approximation is valid both when the support is actually convex and when it is not (Lemmas 2 and 3). This approximation allows us to control the significance level (the probability of type I error), to give approximate values of power (1 minus the probability of type II error, when the significance level is fixed) and to define approximate p-values (the probability that the distribution approximating that of U_n under the null hypothesis gives to values greater than

or equal to the observed value of U_n). The approximate p-value acts as a score of how plausible the support convexity hypothesis is. Our proposal has been empirically validated with a simulation study shown throughout the section. A rigorous proof that the proposed bootstrap-type approximation works (that is, it provides a sequence of probability distributions converging weakly to the same limit distribution as U_n, when n goes to infinity) is beyond the scope of this paper and it probably deserves a separate contribution.

Lemma 2. *Let μ be a probability distribution on $\mathbb{R}^d$ with density f and compact support $S \subset \mathbb{R}^d$. Assume that there exist constants $0 < c < C$ such that $c \leq f(x) \leq C$ for all $x \in S$. Let $X_1, \ldots, X_n$ be i.i.d. vectors drawn from μ. Fix a pair of observations X_i and X_j such that $a = (X_i + X_j)/2$ is in the interior of the support S. Let $h_{(1)}(i,j)$ be defined as above, $h_{(1)}(i,j) = \operatorname{argmin}_h \|X_h - a\|$, and introduce*

$$h_{(2)}(i,j) = \operatorname*{argmin}_{h \neq h_{(1)}(i,j)} \|X_h - a\|.$$

Let $G_{(k)} = \gamma(X_i, X_j, X_{h_{(k)}(i,j)})$, $k = 1, 2$, be the two smallest elements of the ordered sample of $G_h = \gamma(X_i, X_j, X_h)$, $h = 1, \ldots, n$. Assume that f is continuous at a. Then, conditioning on (X_i, X_j), $nG_{(1)}$ and $n(G_{(2)} - G_{(1)})$ converge in distribution, as $n \to \infty$, to an exponential distribution with expected value $(f(a)v_d)^{-1}$, where v_d is the volume of the d-dimensional unit Euclidean ball. Moreover $nG_{(1)}$ and $n(G_{(2)} - G_{(1)})$ are asymptotically independent, given (X_i, X_j).

Proof. Let $D_{(1)} = \|a - X_{h_{(1)}(i,j)}\| = G_{(1)}^{1/d}$. For $0 < s < \|X_i - X_j\|/2$ such that $B(a,s) \subseteq S$,

$$\begin{aligned}
\mathbb{P}\{D_{(1)} > s | X_i, X_j\} &= \mathbb{P}\{\|X_h - a\| > s : h = 1, \ldots, n | X_i, X_j\} \\
&= (1 - \mu(B(a,s)))^{n-2} \\
&= \left(1 - f(a)v_d s^d + o(s^d)\right)^{n-2},
\end{aligned}$$

as s goes to zero where the continuity of f at a is used in the last step. Observe that $f(a) < \infty$ because $a \in S$. Therefore, for $0 < t < n\|X_i - X_j\|^d/2^d$,

$$\begin{aligned}
\mathbb{P}\{nG_{(1)} > t | X_i, X_j\} &= P\left\{nD_{(1)}^d > t | X_i, X_j\right\} \\
&= \mathbb{P}\{D_{(1)} > (t/n)^{1/d} | X_i, X_j\} \\
&= \left(1 - (f(a)v_d t + o(1))\,\frac{1}{n}\right)^{n-2} \\
&\to e^{-f(a)v_d t} \text{ as } n \to \infty,
\end{aligned}$$

and the first part of the lemma is proved.

We prove now the result for $k = 2$. Let $D_{(2)} = \|a - X_{h_{(2)}(i,j)}\| = G_{(2)}^{1/d}$. Then, reasoning as before, for $0 < s + d_1 < \|X_i - X_j\|/2$ such that $B(a, s + d_1) \subseteq S$,

$$\begin{aligned}
&\mathbb{P}\{D_{(2)} > s + d_1 | D_{(1)} = d_1, X_i, X_j\} \\
&\quad = \mathbb{P}\{\|X_h - a\| > s + d_1 : h = 1, \ldots, n, h \neq h_{(1)}(i,j) | X_i, X_j\}
\end{aligned}$$

$$
\begin{aligned}
&= (1-\mu(B(a,s+d_1)\setminus B(a,d_1)))^{n-3} \\
&= \left(1-f(a)v_d\left((s+d_1)^d-d_1^d\right)+o(s^d)\right)^{n-3},
\end{aligned}
$$

as s goes to zero. Then, for $0<t<n\|X_i-X_j\|^d/2^d-nd_1^d$ and defining $d_1=g_1^{1/d}$,

$$
\begin{aligned}
&\mathbb{P}\left\{n(G_{(2)}-G_{(1)})>t|G_{(1)}=g_1,X_i,X_j\right\} \\
&= \mathbb{P}\left\{n(G_{(2)}-d_1^d)>t|D_{(1)}=d_1,X_i,X_j\right\} \\
&= \mathbb{P}\left\{G_{(2)}>d_1^d+t/n|D_{(1)}=d_1,X_i,X_j\right\} \\
&= \mathbb{P}\left\{D_{(2)}^d>d_1^d+t/n|D_{(1)}=d_1,X_i,X_j\right\} \\
&= \mathbb{P}\left\{D_{(2)}>\left(d_1^d+t/n\right)^{1/d}\middle| D_{(1)}=d_1,X_i,X_j\right\} \\
&= \mathbb{P}\left\{D_{(2)}>d_1+\left[\left(d_1^d+t/n\right)^{1/d}-d_1\right]\middle| D_{(1)}=d_1,X_i,X_j\right\} \\
&= (1-f(a)v_dt/n+o(t/n))^{n-3} \\
&= \left(1-(f(a)v_dt+o(1))\frac{1}{n}\right)^{n-3} \\
&\to e^{-f(a)v_dt} \text{ as } n\to\infty,
\end{aligned}
$$

as it was stated. The asymptotic independence between $G_{(2)}$ and $G_{(1)}$ follows from the observation that the conditional distribution of $G_{(2)}$ given that $G_{(1)}=g_1$ does not depend on the value g_1. □

Motivated by Lemma 2, we may define the statistic

$$
U_n^{(2)}=\frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(\gamma(X_i,X_j,X_{h_{(2)}(i,j)})-\gamma(X_i,X_j,X_{h_{(1)}(i,j)})\right).
$$

Lemma 2 establishes that, under the hypothesis of support convexity, each term in the sum defining $U_n^{(2)}$,

$$
G_{(2)}-G_{(1)}=\left(\gamma(X_i,X_j,X_{h_{(2)}(i,j)})-\gamma(X_i,X_j,X_{h_{(1)}(i,j)})\right),
$$

has the same asymptotic conditional distribution as the corresponding term in the sum defining U_n,

$$
G_{(1)}=\gamma(X_i,X_j,X_{h_{(1)}(i,j)}).
$$

Therefore it is reasonable to expect that the asymptotic distributions of U_n and $U_n^{(2)}$ coincide. However, we have not proved that the joint distributions of $\{\gamma(X_i,X_j,X_{h_{(1)}(i,j)}),1\le i<j\le n\}$ and

$$
\left\{\left(\gamma(X_i,X_j,X_{h_{(2)}(i,j)})-\gamma(X_i,X_j,X_{h_{(1)}(i,j)})\right),1\le i<j\le n\right\}
$$

asymptotically coincide, as Lemma 2 states this result only for marginals. This is sufficient to conclude that the expectations of U_n and $U_n^{(2)}$ asymptotically coincide but we cannot make such a statement about the distributions of these statistics.

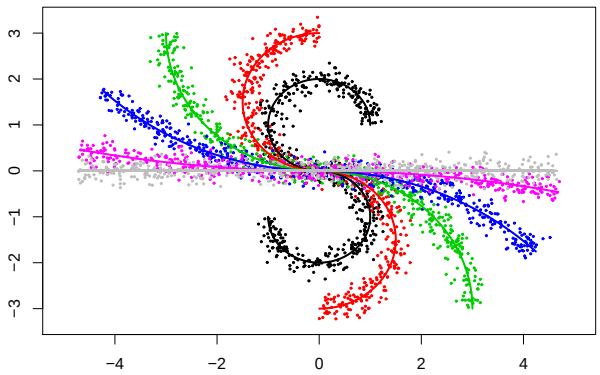

FIG 5. *Example of a convex and 5 non-convex configurations of points: Different two-dimensional S-shaped patterns, with different sharpness. These patterns consist of two circular arches of radius R, with* $R = 1, 1.5, 3, 6, 24, \infty$*, with a constant length equal to* $3\pi/2$*. The bigger the value of R the closer the configuration to convexity, that is achieved for* $R = \infty$*.*

We have carried out some simulations to compile evidence that, under the hypothesis of support convexity, the statistics nU_n and $nU_n^{(2)}$ have similar asymptotic distributions. Consider the different two-dimensional noisy S-shape data sets plotted in Figure 5. They are obtained as follows. Consider two adjacent circumferences both with radius $R > 3/4$. Take an arch of length 1.5π in each of them in such a way that their union forms a differentiable one-dimensional curve of length 3π. Smaller values of the radius R correspond to sharper curves. A flat segment with length 3π corresponds to $R = \infty$. To generate a random point around such a S-shape curve, we generate a random position uniformly distributed over the curve. Then we add an orthogonal deviation from this position, distributed as a truncated normal with zero mean and standard deviation $\sigma = 0.15$, truncated at $[-4\sigma, 4\sigma]$.

Consider now data following a noisy S-shaped pattern with radius $R = \infty$ (that is, a line segment) so that support is a convex set. For sample sizes $n \in \{100, 250, 500, 1000\}$, we have generated 500 data sets. The statistics nU_n and $nU_n^{(2)}$ have been calculated for each sample. The first row of Table 1 shows the p-values of the two-sample Kolmogorov-Smirnov test comparing distributions of nU_n and $nU_n^{(2)}$ (see, for instance, Hollander and Wolfe (1999, Chapter 5)). For large sample sizes the null hypothesis that both statistics have the same distribution can not be rejected. We also have tested the normality of nU_n and $nU_n^{(2)}$ using the Lilliefors normality test (see, for instance, Hollander and Wolfe (1999, Chapter 11)). The corresponding p-values are shown in Table 1. It seems that asymptotic normality is admissible even if U-statistic theory is not directly applicable to these statistics.

In order to establish results under the hypothesis of non-convexity, we need an additional regularity assumption for the support. Given $a \in \mathbb{R}^d \setminus S$ let

$$\pi_S(a) = \{x \in S : \|x - a\| = \min_{y \in S} \|y - a\|\}$$

TABLE 1
p-values of the two-sample Kolmogorov-Smirnov test and the Lilliefors normality test for 500 pairs of observations of statistics nU_n and $nU_n^{(2)}$

Test	Statistics	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
2-sample KS	U_n and $U_n^{(2)}$	<1e-10	.0038	.4131	.0587
Normality	U_n	<1e-10	.0504	.1529	.1729
Normality	$U_n^{(2)}$	<1e-4	.0676	.0010	.6275

be the set of closest points in S to a. When $\pi_S(a)$ has a unique point, we call it a_S. Erdős (1945) proved that the set of points $a \in \mathbb{R}^d - S$ with more than one point in $\pi_S(a)$ has null Lebesgue measure. The required regularity condition is as follows:

(A) For any $a \in \mathbb{R}^d \setminus S$ such that $\pi_S(a) = \{a_S\}$ there exist constants $\eta > 0$ and $\nu \geq 1$, both depending on a and S, such that, when $s \to 0$, $s > 0$,

$$\text{Vol}\left(B(a, \|a - a_S\| + s) \cap S\right) = \eta s^\nu + o(s^\nu).$$

Condition **(A)** is satisfied for many regular sets S, convex or not. For instance, let $S = [-1, 1]^3 \subset \mathbb{R}^3$ and $a = (1 + \delta, 0, 0)$. Then $a_S = (1, 0, 0)$, $\|a - a_S\| = \delta$ and, for small δ and s (such that $s + \delta \leq \|a - (1, 1, 0)\|$), the solid $B(a, \delta + s) \cap S$ is a spherical cap (portion of a sphere cut off by a plane) and its volume is (see, e.g., Li (2011))

$$\text{Vol}\left(B(a, \delta + s) \cap S\right) = \frac{\pi}{3} s^2 (3\delta + 2s) = \pi\delta s^2 + o(s^2),$$

and **(A)** is verified with $\eta = \pi\delta$ and $\nu = 2$. For the volume $V_d(\delta, s)$ of the corresponding d-dimensional hyperspherical cap, Li (2011) gives the following formula:

$$V_d(\delta, s) = \frac{1}{2} v_d (\delta + s)^d I_{1-\{\delta/(\delta+s)\}^2}\left(\frac{d+1}{2}, \frac{1}{2}\right),$$

where $I_x(a, b)$ is the regularized incomplete beta function, defined for $0 \leq x \leq 1$, $a > 0$, $b > 0$, as

$$I_x(a, b) = \frac{1}{B(a, b)} \int_0^x t^{a-1}(1 - t)^{b-1} dt,$$

where $B(a, b) = \int_0^1 t^{a-1}(1 - t)^{b-1} dt$ is the beta function. It is easy to check that $I_x(a, b) = x^a/(aB(a, b)) + o(x^a)$, when $a \to 0$, and that $1 - \{\delta/(\delta + s)\}^2 = 2s/\delta + o(s)$, when $s \to 0$. Therefore, when $s \to 0$,

$$\begin{aligned} V_d(\delta, s) &= \frac{1}{2} v_d (\delta + s)^d \left(\frac{2(2/\delta)^{(d+1)/2}}{(d+1)B((d+1)/2, 1/2)} s^{(d+1)/2} + o(s^{(d+1)/2}) \right) \\ &= \frac{2 v_d (2\delta)^{(d-1)/2}}{(d+1)B((d+1)/2, 1/2)} s^{(d+1)/2} + o(s^{(d+1)/2}). \end{aligned}$$

So $\nu = (d + 1)/2$ in this case.

In the following example in $\mathbb{R}^2$ the value of η depends on the shape of S. Consider $S = \{(x, y) \in \mathbb{R}^2 : 0 \le x \le 1, -x^\alpha \le y \le x^\alpha\}$, for $\alpha > 1/2$. For $\delta > 0$, let $a = (-\delta, 0)$. Then $a_S = (0, 0)$, $\|a - a_S\| = \delta$ and

$$\frac{2}{\alpha+1} s_0^{\alpha+1} \le \mathrm{Vol}\,(B(a, \delta + s) \cap S) \le \frac{2}{\alpha+1} s^{\alpha+1},$$

where $s_0 > 0$ is such that $(\delta + s)^2 = (\delta + s_0)^2 + s_0^{2\alpha}$. For $\alpha > 1/2$ it can be proved that $\lim_{s\to 0}(s/s_0) = 1$. Then

$$\mathrm{Vol}\,(B(a, \delta + s) \cap S) = \frac{2}{\alpha+1} s^{\alpha+1} + o(s^{\alpha+1}).$$

Now we state the result analogous to Lemma 2 when the support S is not convex and the middle point $a = (X_i + X_j)/2$ is not in S.

Lemma 3. *Let μ be a probability distribution on $\mathbb{R}^d$ with density f and compact support $S \subset \mathbb{R}^d$. Assume that there exist constants $0 < c < C$ such that $c \le f(x) \le C$ for all $x \in S$. Let $X_1, \ldots, X_n$ be i.i.d. vectors drawn from μ. Assume that S is not convex and fix a pair of observations X_i and X_j such that $a = (X_i + X_j)/2 \notin S$. Then, with probability one, there is only one closest point in S to a. Let a_S be such a point and assume that f is continuous at a_S.*

Assume additionally that condition **(A)** *is verified. Let $G_{(1)}$ and $G_{(2)}$ be defined as in Lemma 2. Then, conditioning on (X_i, X_j), as $n \to \infty$, $G_{(1)}$ converges in probability to $\|a - a_S\|^d$ and $n(G_{(2)} - G_{(1)})$ converges in distribution to an exponential distribution with expected value*

$$\frac{d\|a - a_S\|^{d-\nu}}{f(a_S)\eta\nu}.$$

Proof. Given Erdős' result on the null Lebesgue measure of the set of points with more than one closest point in S, this set has zero probability because $(X_i + X_j)/2$ is absolutely continuous.

Now, let $D_{(1)} = \|a - X_{h_{(1)}(i,j)}\| = G_{(1)}^{1/d}$ and $s > 0$. Then, arguing as in the proof of Lemma 2 and using **(A)**,

$$\begin{aligned}
\mathbb{P}\{|D_{(1)} - \|a - a_S\|| > s\} &= \mathbb{P}\{D_{(1)} > \|a - a_S\| + s\} \\
&= (1 - \mu(B(a, \|a - a_S\| + s) \cap S))^n \\
&= (1 - f(a_S)\eta s^\nu + o(s^\nu))^n
\end{aligned}$$

as s goes to zero. Therefore, for $t > 0$

$$\begin{aligned}
\mathbb{P}\{n^{1/\nu}(D_{(1)} - \|a - a_S\|) > t\} &= \mathbb{P}\left\{D_{(1)} > \|a - a_S\| + t/n^{1/\nu}\right\} \\
&= \left(1 - f(a_S)\eta\left(t/n^{1/\nu}\right)^\nu + o(1/n)\right)^n \\
&\to e^{-f(a_S)\eta t^\nu} \text{ as } n \to \infty.
\end{aligned}$$

It follows that $n^{1/\nu}(D_{(1)} - \|a - a_S\|)$ converges in distribution to a Weibull distribution with shape parameter ν. It follows that $D_{(1)}$ converges in probability to $\|a - a_S\|$ and, by continuity of $g(x) = x^{d/\nu}$, that $G_{(1)}$ converges in probability to $\|a - a_S\|^d$ and the first part of the lemma is proved.

Defining $D_{(2)} = \|a - X_{h_{(2)}(i,j)}\| = G_{(2)}^{1/d}$, proceeding again as in the proof of Lemma 2 and using **(A)**, for $s > 0$

$$\begin{aligned}
&\mathbb{P}\{D_{(2)} > s + d_1 | D_{(1)} = d_1\} \\
&\quad = \quad (1 - \mu\{(B(a, s + d_1) \setminus B(a, d_1)) \cap S\})^{n-1} \\
&\quad = \quad (1 - f(a_S)\eta(\,(s + d_1)^\nu - d_1^\nu\,) + o(s^\nu))^{n-1}
\end{aligned}$$

as s goes to zero. Then, for $t > 0$

$$\begin{aligned}
&\mathbb{P}\left\{n(D_{(2)}^\nu - D_{(1)}^\nu) > t | D_{(1)} = d_1\right\} \\
&\quad = \quad \mathbb{P}\left\{D_{(2)}^\nu > d_1^\nu + t/n | D_{(1)} = d_1\right\} \\
&\quad = \quad \mathbb{P}\left\{D_{(2)} > \{(d_1^\nu + t/n)^{1/\nu} - d_1\} + d_1 | D_{(1)} = d_1\right\} \\
&\quad = \quad (1 - f(a_S)\eta t/n + o(1/n))^{n-1} \\
&\quad \to \quad e^{-f(a_S)\eta t} \text{ as } n \to \infty.
\end{aligned}$$

In the last equality we have used that

$$s = \left((d_1^\nu + t/n)^{1/\nu} - d_1\right) = \frac{1}{\nu} d_1^{1-1/\nu} t/n + o(1/n)$$

as n goes to ∞. Then

$$s^\nu = \left((d_1^\nu + t/n)^{1/\nu} - d_1\right)^\nu = O(1/n^\nu),$$

as n goes to ∞. Therefore $o(s^\nu) = o(1/n)$ because $\nu \geq 1$.

It follows that $n(D_{(2)}^\nu - D_{(1)}^\nu)$, given $D_{(1)} = d_1$, converges in distribution to D, an exponential random variable with expectation $(f(a_S)\eta)^{-1}$. Observe that $n(D_{(2)}^\nu - D_{(1)}^\nu)$ and $D_{(1)}$ are asymptotically independent, given (X_i, X_j). It follows that $n(D_{(2)}^\nu - D_{(1)}^\nu)$ also converges to D without conditioning on $D_{(1)}$.

To prove the asymptotic distribution of $n(G_{(2)} - G_{(1)})$ we need a minor variation of the proof of Cramér Delta Theorem provided in Arnold (1990). Consider $g(x) = x^{d/\nu}$. Then $n(G_{(2)} - G_{(1)}) = n(g(D_{(2)}^\nu) - g(D_{(1)}^\nu))$. By Taylor's theorem,

$$\begin{aligned}
&n(G_{(2)} - G_{(1)}) \\
&= n\left\{g(D_{(1)}^\nu) + g'(D_{(1)}^\nu)(D_{(2)}^\nu - D_{(1)}^\nu) + (D_{(2)}^\nu - D_{(1)}^\nu) R(D_{(2)}^\nu - D_{(1)}^\nu) - g(D_{(1)}^\nu)\right\} \\
&= g'(D_{(1)}^\nu)\left\{n(D_{(2)}^\nu - D_{(1)}^\nu)\right\} + \left\{n(D_{(2)}^\nu - D_{(1)}^\nu)\right\} R(D_{(2)}^\nu - D_{(1)}^\nu),
\end{aligned}$$

where $R(x) \to R(0) = 0$ as $x \to 0$ (and then $R(x)$ is continuous at 0). Observe that

$$D_{(2)}^{\nu} - D_{(1)}^{\nu} = \frac{1}{n}\left(n(D_{(2)}^{\nu} - D_{(1)}^{\nu})\right) \to 0 \text{ in probability as } n \to \infty$$

because $n(D_{(2)}^{\nu} - D_{(1)}^{\nu}) \to D$ weakly. Then the continuity of $R(x)$ at 0 and Slutzky's theorem (see, for instance, Arnold 1990, Theorem 6.8) give

$$\left\{n(D_{(2)}^{\nu} - D_{(1)}^{\nu})\right\} R(D_{(2)}^{\nu} - D_{(2)}^{\nu}) \to 0$$

in probability as $n \to \infty$. Moreover $g'(D_{(1)}^{\nu}) \to g'(\|a - a_S\|^{\nu})$ by the continuity of $g'(x) = (d/\nu)x^{(d/\nu)-1}$ at $(0, \infty)$. It follows (again by the Slutzky's theorem) that $n(G_{(2)} - G_{(1)})$ has the same limit distribution as $g'(\|a - a_S\|^{\nu})n(D_{(2)}^{\nu} - D_{(1)}^{\nu})$, but

$$g'(\|a - a_S\|^{\nu})n(D_{(2)}^{\nu} - D_{(1)}^{\nu}) \to g'(\|a - a_S\|^{\nu})D = \frac{d}{\nu}\|a - a_S\|^{d-\nu}D \text{ weakly,}$$

the limit distribution being an exponential distribution with expected value $d\|a - a_S\|^{d-\nu}(f(a_S)\eta\nu)^{-1}$, and the proof concludes. □

It follows from Lemma 3 that, given (X_i, X_j) with $a = (X_i + X_j)/2 \notin S$,

$$\begin{aligned} nG_{(1)} &= nD_{(1)}^{d} \\ &= \left\{n^{\frac{1}{d}-\frac{1}{\nu}}\left[n^{\frac{1}{\nu}}(D_{(1)} - \|a - a_S\|)\right] + n^{\frac{1}{d}}\|a - a_S\|\right\}^{d} \\ &= \left\{n^{\frac{1}{d}-\frac{1}{\nu}}O_p(1) + n^{\frac{1}{d}}\|a - a_S\|\right\}^{d} \\ &= n\|a - a_S\|^{d} + O_p(n^{\frac{1}{d}-\frac{1}{\nu}}). \end{aligned}$$

Therefore, $nG_{(1)}$ goes to infinity (in probability) at rate n.

Lemma 3 suggests that for a non-convex support S and assuming **(A)**, $nU_n^{(2)}$ should be bounded in probability (because $nU_n^{(2)}$ is the average of $n(n-1)/2$ random variables that are bounded in probability) but nU_n should not be (because it is the average of $n(n-1)/2$ random variables that go to infinity at rate n). In fact, from Proposition 1 it follows that $\lim_n nU_n = \infty$ almost surely when the support S is not convex.

To understand the intuitive meaning of Lemmas 2 and 3, let $F_{nU_n}^{\text{conv}}$ and $F_{nU_n^{(2)}}^{\text{conv}}$ be the distributions of the two statistics under consideration, nU_n and $nU_n^{(2)}$, respectively, when the support S is convex. For the case of non-convex support, we call $F_{nU_n}^{\text{non-conv}}$ and $F_{nU_n^{(2)}}^{\text{non-conv}}$ the distributions of the two corresponding statistics. Lemma 2 establishes that $F_{nU_n}^{\text{conv}}$ and $F_{nU_n^{(2)}}^{\text{conv}}$ are similar and Lemma 3 indicates that $F_{nU_n^{(2)}}^{\text{non-conv}}$ looks more like $F_{nU_n}^{\text{conv}}$ than $F_{nU_n}^{\text{non-conv}}$. Therefore we propose to use the distribution of the statistic $nU_n^{(2)}$ to approximate that of nU_n under the support convexity hypothesis, whether the support is indeed convex or not.

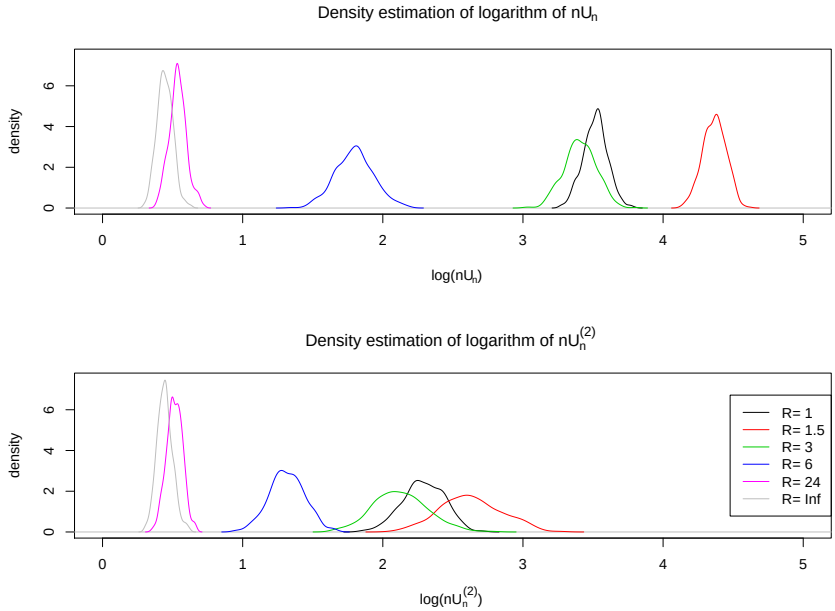

FIG 6. *Up: Density estimation of* 500 *observed values of the statistic* $\log nU_n$ *for each of the six noisy S-shaped configurations (see Figure 5) for sample size* $n = 500$. *Down: Densities of* 500 *values of the statistic* $\log nU_n^{(2)}$ *generated from data according to each of the six noisy S-shaped configurations for sample size* $n = 500$.

Some simulation have been conducted to support the use of the distribution of $U_n^{(2)}$ to approximate that of U_n. The upper panel of Figure 6 shows the estimated density of the statistic $\log nU_n$ calculated on 500 samples (of size $n = 500$) generated according to each of the six S-shaped supports shown in Figure 5. It can be clearly seen how the distribution of nU_n changes in a considerable way with the data pattern and how its values get closer to 0 as the support gets closer to convexity.

The lower panel of Figure 6 shows estimated densities (from 500 values) of $\log nU_n^{(2)}$ calculated over 500 samples (of size $n = 500$) generated according to each of the six S-shaped configurations shown in Figure 5. The scale of upper panel has been kept in order to clearly show how the distributions of $\log(nU_n^{(2)})$ under non-convexity are closer to 0 than those of $\log(nU_n)$. This is the main conclusion of the simulations.

Observe that most of the estimated densities of $\log(nU_n^{(2)})$ corresponding to non-convex S do not overlap the density of $\log(nU_n)$ corresponding to convex S (i.e., $R = \infty$). This fact does not necessarily contradict our belief that the null distribution of nU_n can always be approximated with that of $nU_n^{(2)}$: the asymptotic distribution of $n(G_{(2)} - G_{(1)})$ depends on whether the data follow the null distribution or not (see Lemmas 2 and 3).

TABLE 2
Empirical significance levels for three nominal significance levels α. Those being significantly different from the nominal ones (95% confidence) have been written in italics

α	$n = 100$	$n = 250$	$n = 500$	$n = 1000$
.01	.006	.004	.016	.012
.05	*.022*	*.020*	.068	.058
.1	*.026*	*.044*	*.134*	*.134*

The use of the distribution of the statistic $nU_n^{(2)}$ as an approximation for that of nU_n under the null hypothesis entails a problem: only one observation of $nU_n^{(2)}$ is available from each sample $X_1, \ldots, X_n$. Therefore a resampling procedure is required to provide a set of pseudo-observations of $nU_n^{(2)}$.

The standard bootstrap is not adequate in this context because at each bootstrap sample there would be some repeated observations, say $X_i^* = X_j^* = X_l$, for $i \neq j$, and therefore we would have $\min_{h=1\ldots n} \gamma(X_i^*, X_j^*, X_h^*) = 0$, thus reducing the effective number of summands defining U_n^*.

We propose to perform subsampling bootstrap, that is, resampling without replacement from the original sample at smaller than the original sample size (see Politis and Romano (1994); Politis, Romano and Wolf (1999)). Then the procedure to compute p-values for the support convexity decision rule is as follows. Let nU_n^{Obs} be the observed values of nU_n for the sample $X_1, \ldots, X_n$ at hand. We take B subsamples of size $m < n$ and compute the statistic $mU_m^{(2)}$ for each subsample. Let $mU_{m,b}^{(2)*}$, $b = 1, \ldots, B$, be the B values of $mU_m^{(2)}$ obtained this way. Let μ_m^* and s_m^* be the sample mean and standard deviation, respectively, of $mU_{m,b}^{(2)*}$, $b = 1, \ldots, B$. We approximate the distribution of nU_n (under support convexity) by a normal distribution centered at μ_m^* and having standard deviation equal to s_m^*. Let Φ be the distribution function of the standard normal distribution. The p-value is therefore defined as

$$p - \text{value} = 1 - \Phi\left(\frac{nU_n^{\text{Obs}} - \mu_m^*}{s_m^*}\right). \tag{4}$$

Table 2 illustrates the performance of the proposed procedure for deciding about support convexity when this is true. For each different sample size, 500 samples have been generated according to the hypothesis of support convexity (S-shape pattern with $R = \infty$). For each sample, $B = 100$ bootstrap subsamples have been drawn, with sizes $m = n/2$ (for $n \in \{100, 250, 500\}$) or $m = n/4$ (for $n = 1000$). The empirical significance levels are calculated as the proportion of samples for which the computed p-value is lower than the nominal one. We see that the nominal significance level is well reproduced for $\alpha = 0.01$ and $\alpha = 0.05$ (when $n \geq 500$ in this case), but the case $\alpha = 0.1$ is unsatisfactory.

The empirical power of the convexity decision rule has been calculated (for a nominal significance value $\alpha = 0.05$) for sample sizes $n \in \{100, 250, 500, 1000\}$ as the proportion of samples for which the computed p-value is lower than α. Figure 7 shows the estimated powers. It can be seen how patterns which are distant from convexity are perfectly detected. For a sample size of $n = 100$ the

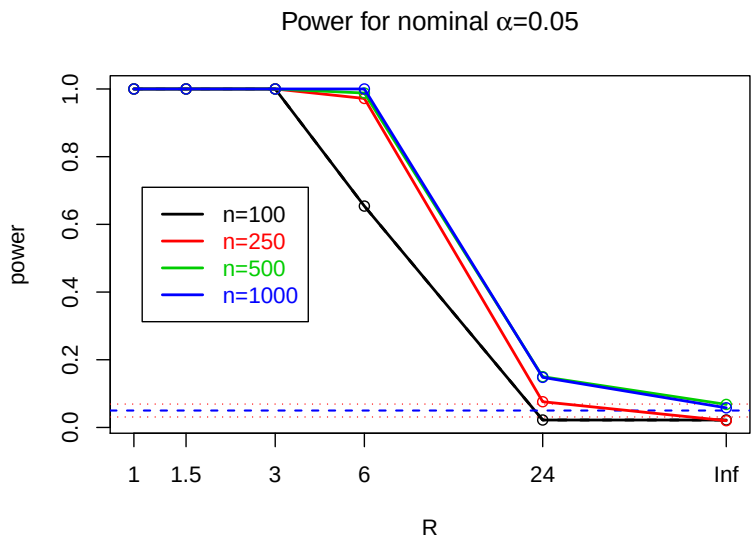

FIG 7. *Empirical power functions (for nominal significance level* $\alpha = 0.05$*) estimated from 500 samples (*$n \in \{100, 250, 500, 1000\}$*). The parameter* R *indicates the radius of the circumferences used to produce the six noisy S-shaped configurations in Figure 5: the bigger the value of* R*, the closer the support is to convexity, which is achieved at* $R = \infty$*. Horizontal lines mark acceptance intervals of the null hypothesis that the observed powers equal to the nominal significance level* $\alpha = 0.05$*.*

case closest to convexity ($R = 24$) is not detected as non-convex while it is detected when $n \in \{250, 500, 1000\}$.

5. Choice of the tuning parameter in ISOMAP

In this section we present a statistical application of the rule introduced in Section 2. We use this decision rule for choosing automatically the tuning parameter of ISOMAP, a nonlinear dimensionality reduction method due to Tenenbaum, de Silva and Langford (2000).

Given n points $x_1, \ldots, x_n \in \mathbb{R}^p$ in a high-dimensional space, equipped with metric d, the object of nonlinear dimensionality reduction (also known as manifold learning) is to find a low dimensional configuration, that is, an $n \times d$ matrix, with $d \ll p$, with rows y_i, $i = 1, \ldots, n$, and a nonlinear function $\rho : \mathbb{R}^d \to \mathbb{R}^p$ such that $\rho(y_i)$ is close (in some sense) to the observed x_i, for $i = 1, \ldots, n$. Principal Component Analysis (PCA) is without doubt the most used dimensionality reduction technique, but it is not able to detect nonlinear structures. See Lee and Verleysen (2007) or Gorban et al. (2007) for a broad coverage of nonlinear dimensionality reduction.

We focus on ISOMAP transformation. The underlying implicit assumption is that the high-dimensional data lie on, or close to, a low-dimensional nonlinear manifold and the geodesic distance of the manifold represents a meaningful metric. ISOMAP tries to recover this hidden information.

The algorithm takes as its starting point the distance matrix $D = (d(x_i, x_j))_{i,j}$ between all pairs of points in the original space of high dimension. The available distance $d(x_i, x_j)$ is a good approximation of the geodesic distances only for those pairs of points x_i and x_j that are *close enough*.

The ISOMAP algorithm can be briefly described as a three-step process:

1. First step determines that points x_i and x_j are *neighbors* in the original space if $d(x_i, x_j) \leq \epsilon$, for a $\epsilon > 0$. (Another version declares that two points are neighbors if one of them is one of the k nearest neighbors of the other). A weighted graph G_ϵ is built with vertices being the data points and edges of weight $d(x_i, x_j)$ between neighboring points.
2. In the second step, the matrix of geodesic distances $d_G^\epsilon(x_i, x_j)$ between all pairs of points is estimated by computing shortest path distances in the graph G_ϵ. The matrix $D_G^\epsilon = (d_G^\epsilon(x_i, x_j))_{i,j}$ is obtained.
3. In the third step, classical multidimensional scaling (MDS) is applied to the matrix of distances D_G^ϵ in order to obtain coordinates for the points in $\mathbb{R}^d$, with $d < p$.

For the sake of completeness, we recall that MDS is the generic name of a variety of techniques for the analysis of dissimilarities (or distances) on a set of n objects. The main objective of MDS (see, for instance, Borg and Groenen (2005) for more details) is to provide a representation of the objects as points in a geometric space in such a way that the matrix of Euclidean distances in this space is as close as possible to the dissimilarity matrix. One of the first models used for MDS was classical MDS (Borg and Groenen (2005, Chapter 12)). It is based on the assumption that a representation exists such that the Euclidean distance equals the dissimilarity matrix. Then it is possible to find this Euclidean representation algebraically without an iterative algorithm.

One of the advantages of the ISOMAP algorithm is that only one parameter, ϵ, is required in step 1. However, the performance of the algorithm crucially depends on the choice of this parameter. If the neighborhood is too large, local neighbors will include data points from other branches of the manifold and *short-circuits* will appear: faraway points according to geodesic distances would turn up to be close according to the Euclidean distance. On the other hand, if the neighborhood is too small, the manifold will fragment into disconnected clusters and the algorithm will not be able to assign distances between every pair of points.

In the original work (Tenenbaum, de Silva and Langford (2000)) this parameter was chosen manually. There have been other methods for selecting the optimal parameter value. See Shao, Huang and Wan (2007) and references therein.

Our proposal for the selection of the tuning parameter ϵ is motivated by the following observation. The original distance $d(x_i, x_j)$ between points x_i and x_j is similar to their geodesic distance $d_G^\epsilon(x_i, x_j)$ when either x_i and x_j are within distance ϵ or when x_i and x_j can be connected through a path that contains points $h_0 = x_i, h_1, \ldots, h_r, h_{r+1} = x_j$ in such a way that for all $k = 0, \ldots, r$, h_k

and h_{k+1} are within distance ϵ and also

$$\sum_{k=0}^{r} d(h_k, h_{k+1}) \approx d(x_i, x_j). \tag{5}$$

If this happens for all pairs of points, there is no need to modify the distance matrix (steps 1 and 2 of ISOMAP) and MDS can be applied directly. Therefore, the full ISOMAP algorithm should be applied with a tuning parameter ϵ such that in the final configuration a condition similar to (5) holds:

$$\sum_{k=0}^{r} d_G^{\epsilon}(h_k, h_{k+1}) \approx d_G^{\epsilon}(x_i, x_j). \tag{6}$$

Note that in the space $\mathbb{R}^d$ only Euclidean distance is applied, and also that condition (6) for Euclidean spaces is equivalent to stating that around the segment $[x_i, x_j]$ joining points x_i and x_j, there are other sample points, that is, this segment crosses completely the support of the underlying probability distribution. In other words, roughly, the Euclidean distance between x_i and x_j is substantially different from their geodesic distance if and only if the segment $[x_i, x_j]$ is not completely included in the support of the distribution. Therefore, it is possible to find pairs of points (x_i, x_j) with this property if and only if the support of the underlying probability distribution is not a convex set. As a consequence, the tuning parameter in ISOMAP should be chosen in order to guarantee that the final configuration is compatible with an underlying probability distribution with a convex support.

Based on the intuitive argument described above, our proposal is based on applying to distance matrix D_G^{ϵ} the support convexity decision rule. This decision rule is used to assign, to any possible value of ϵ, a score according to the plausibility of the support convexity hypothesis for the corresponding configuration generated by ISOMAP. The selected parameter is the one achieving the maximum score. We implicitly assume that the embedded low dimensional distribution has convex support and that the possible non-convexity in the high dimensional space is due only to the embedding being non-linear. Our proposal would not work in situations as, for instance, that of a two-dimensional distribution whose support is a circle with two holes that is non-linearly embedded in a three-dimensional space.

Let $(\epsilon_{\min}, \epsilon_{\max})$ be an interval of candidate values for parameter ϵ. Given $\epsilon \in (\epsilon_{\min}, \epsilon_{\max})$, the ϵ-ISOMAP algorithm is applied to D, the starting distance matrix. Let D_G^{ϵ} be the output distance matrix (that is, the matrix containing Euclidean distances between points in the low-dimensional space). Then the support convexity decision rule is performed from D_G^{ϵ}. Let $p(\epsilon)$ be the p-value defined in (4). We use $p(\epsilon)$ as the score value for ϵ.

We propose to choose parameter ϵ as

$$\epsilon^* = \arg\max\{p(\epsilon) : \epsilon \in (\epsilon_{\min}, \epsilon_{\max})\}.$$

Therefore ϵ^* is the value in $(\epsilon_{\min}, \epsilon_{\max})$ for which the highest compatibility with the hypothesis of support convexity is achieved. A remark on the meaningful

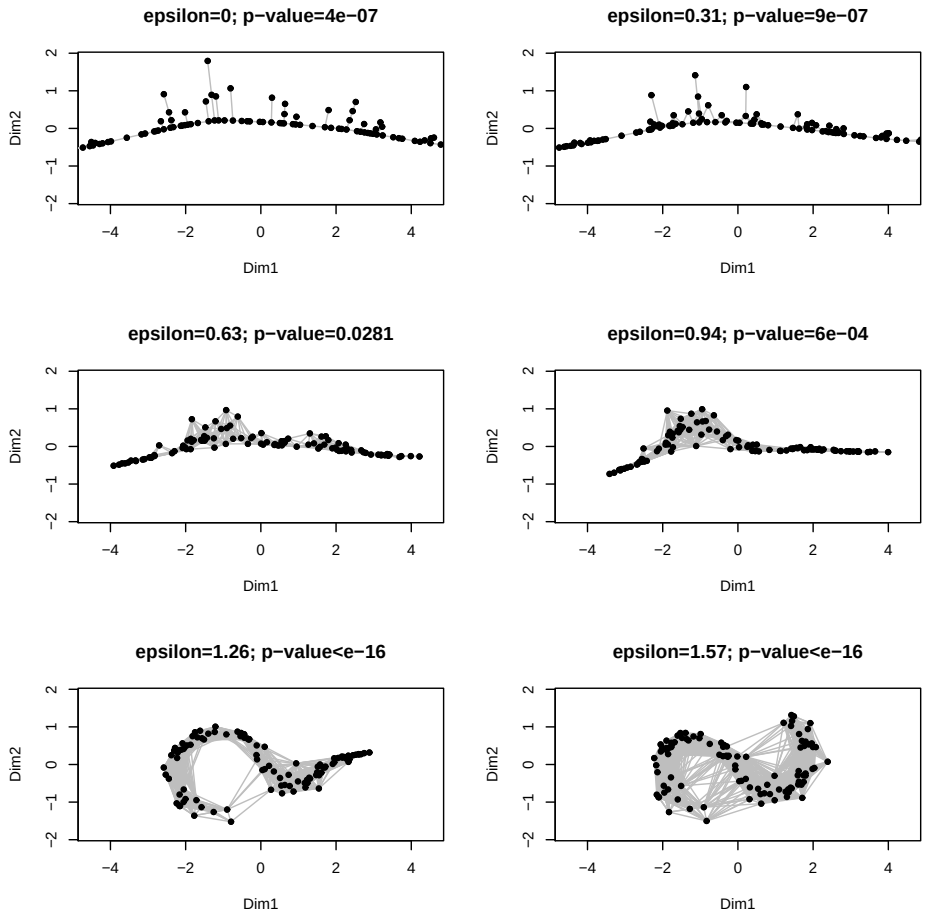

FIG 8. *Choosing ϵ for the sharpest S-shaped configuration (see Figure 5) with sample size $n = 100$. Low dimensional configurations obtained as output of the ϵ-ISOMAP algorithm for six values of ϵ, and the corresponding p-values obtained when deciding about support convexity.*

choice of $\epsilon_{\min}$ and $\epsilon_{\max}$ can be found in the Appendix. There you can also find a proposal to avoid disconnected graphs (even if we take $\epsilon = 0$).

We have applied the proposed procedure for the choice of ϵ to a bi-dimensional synthetic data set, corresponding to the sharpest S in Figure 5, that with radius $R = 1$. The sample size is $n = 100$ and $B = 200$ bootstrap samples are obtained. We have used values ϵ equal to seven evenly spaced values $\epsilon_1, \ldots, \epsilon_7$, with $\epsilon_1 = 0$ and $\epsilon_7 = \text{median}\{d(x_i, x_j)\}$. The resulting p-values are shown in Figure 8 (only for ϵ_1 to ϵ_6; the result for ϵ_7 is very similar to that of ϵ_6).

The two panels in the first row of Figure 8 show that when ϵ is too small, a linear representation is adequate for many points in the sample, but there appear some points that are too far from the common linear structure, making the support convexity hypothesis hard to be accepted. For moderate values of ϵ

(second row of Figure 8) a compromise is achieved between a common linear structure and the absence of outliers. When $\epsilon = 1.26$ (lower left panel) a short-circuit appears (leading to a misleading bi-dimensional configuration). It could be said that the short-circuit is also present (but not so clearly) for $\epsilon = 0.94$. For $\epsilon = 1.57$ (lower right panel) there are two short-circuits. Following our proposal, the chosen value for ϵ is $\epsilon^* = 0.63$. Observe that the p-value corresponding to this best choice of ϵ is 0.0281, indicating a moderate evidence against the null hypothesis of support convexity. This happens because the ISOMAP procedure is not able to fully linearize the data configuration even for the most favorable value of parameter ϵ.

6. Conclusions

In this paper we investigated the possibilities and limitations of constructing data-based procedures to decide whether the support of the underlying density generating the data points is convex or nor. We defined a decision rule, based on a U-statistic with a random kernel, which decides correctly for sufficiently large n, with probability 1, whenever the density is bounded away from zero in its compact support and the support has a boundary of zero Lebesgue measure.

We also show that such asymptotically correct decision rules are impossible to define if one only assumes boundedness of the density.

Moreover, we suggest a bootstrap-like procedure for approximating the distribution of the proposed test statistic under the hypothesis of convexity of the support. The performance of the proposed method is illustrated on simulated data sets.

To illustrate potential applications, the decision rule is used to automatically choose the tuning parameter of ISOMAP, a nonlinear dimensionality reduction method.

Acknowledgments

We thank the referees for their important remarks that helped improve the presentation.

Appendix

Some simple lemmas

Lemma 4. *Let f be a density on $\mathbb{R}^d$ with support S and assume that there exists $c > 0$ such that $f(x) \geq c$ for all $x \in S$. Then* $\mathrm{Vol}(\partial S) = 0$ *if and only if for almost every $x \in S$ there exists $\epsilon > 0$ such that* $\operatorname{essinf}_{y:\|y-x\|<\epsilon} f(y) > 0$.

Proof. First note that every $x \in S \setminus \partial S$ is an interior point of S. Since f is bounded away from zero on its support, for all such points there exists $\epsilon > 0$ such that $\operatorname{essinf}_{y:\|y-x\|<\epsilon} f(y) \geq c$. Thus if $\mathrm{Vol}(\partial S) = 0$, the Lebesgue measure of $x \in S$ with $\operatorname{essinf}_{y:\|y-x\|<\epsilon} f(y) = 0$ for all $\epsilon > 0$ equals zero.

On the other hand, for every $x \in \partial S$, $\operatorname{essinf}_{y:\|y-x\|<\epsilon} f(y) = 0$ for all $\epsilon > 0$. To see this, suppose that this is not true and for some $x \in \partial S$, there exists $\epsilon > 0$ such that $\operatorname{essinf}_{y:\|y-x\|<\epsilon} f(y) > 0$. But since x is on the boundary of S, there exists $z \notin S$ such that $\|z - x\| < \epsilon/2$. Since S is closed, there exists $\delta < \epsilon/2$ such that the ball $N(z, \delta)$ is entirely outside of S. But then $\operatorname{essinf}_{y \in N(z,\delta)} f(y) > 0$, which contradicts the definition of the support.

This implies that if $\mathrm{Vol}(\partial S) > 0$, then the Lebesgue measure of the set of points $x \in S$ with $\operatorname{essinf}_{y:\|y-x\|<\epsilon} f(y) = 0$ for all $\epsilon > 0$ is positive. □

Lemma 5. *Let f be a density with support S. Suppose that* $\mathrm{Vol}(\partial S) = 0$ *and $f(x) \geq c$ for all $x \in S$ where $c > 0$. Define the set $A = \{x : \exists \delta > 0 : \operatorname{essinf}_{y \in N(x,\delta)} f(y) > 0\}$. Then S is the closure of A.*

Proof. Since $A \subset S$ and S is closed, $\overline{A} \subseteq S$ (where $\overline{A}$ stands for the closure of A). Suppose $S \neq \overline{A}$. Then there exists $x \in S$ and $\epsilon > 0$ such that $N(x, \epsilon) \cap A = \emptyset$. Observe that since $f(x) \geq c$ on S, for every point $y \in N(x, \epsilon)$, either $y \notin S$ or $y \in \partial S$. Thus, by assumption, $\mathrm{Vol}(S \cap N(x, \epsilon)) = 0$. But then the closed set $S \cap N(x, \epsilon)^c$ has f-measure 1 which contradicts the definition of S (since the support is defined as the smallest closed set of f-measure 1). □

Some technical details on the choice of ϵ^*

A remark on the meaningful choice of $\epsilon_{\min}$ and $\epsilon_{\max}$ follows. Very low values of ϵ produce disconnected graphs G_ϵ in the first step of the ϵ-ISOMAP algorithm. Then the usual way to circumvent the problem is to analyze only the largest connected component of G_ϵ. Then different samples are used for different values $\epsilon < \epsilon_{\mathrm{conn}}$, being that value the lowest one assuring the connectivity of G_ϵ. So it may seem plausible to take $\epsilon_{\min} = \epsilon_{\mathrm{conn}}$. Unfortunately, the value of ϵ_{conn} is extremely sensitive to outliers, because $\epsilon_{\mathrm{conn}} \geq \max_i \min_j d(x_i, x_j)$.

Our proposal to avoid disconnected graphs G_ϵ for small ϵ is based on the Minimum Spanning Tree associated to distance matrix D. Let G^0_{MST} be the graph representing this Minimum Spanning Tree, which is connected by definition. We propose to replace always the graph G in the first step of ϵ-ISOMAP algorithm by the union graph $G^\epsilon_{\mathrm{MST}} = G_\epsilon \cup G^0_{\mathrm{MST}}$, and proceed to further steps in the usual way. Observe that G_{MST} is connected for all $\epsilon \geq 0$, being $G_{\mathrm{MST}} = G^0_{\mathrm{MST}}$ for $\epsilon = 0$. Therefore we may choose $\epsilon_{\min} = 0$.

An easy way to fix $\epsilon_{\max}$ is taking

$$\epsilon_{\max} = \max_{i,j} d(x_i, x_j).$$

This choice allows one the possibility of having observed a distance matrix D compatible with a convex support probability distribution. In practice a lower value may be chosen such as $\epsilon_{\max} = \mathrm{median}\{d(x_i, x_j)\}$. Then a fine regular grid $\epsilon_1 = \epsilon_{\min} < \cdots < \epsilon_E = \epsilon_{\max}$ is used and p-values are computed: $p(\epsilon_e), e = 1, \ldots, E$.

References

Arnold, S. F. (1990). *Mathematical Statistics*. Prentice Hall.

Baíllo, A., Cuevas, A. and Justel, A. (2000). Set estimation and nonparametric detection. *The Canadian Journal of Statistics* **28** 765–782. MR1821433

Baíllo, A. and Cuevas, A. (2001). On the estimation of a star-shaped set. *Advances in Applied Probability* **33** 717–726. MR1875774

Biau, G., Cadre, B. and Pelletier, B. (2008). Exact rates in density support estimation. *Journal of Multivariate Analysis* **99** 2185–2207. MR2463383

Borg, I. and Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications (2nd ed)*. Springer-Verlag, New York. MR2158691

Boucheron, S., Bousquet, O., Lugosi, G. and Massart, P. (2005). Moment inequalities for functions of independent random variables. *The Annals Probability* **33** 514–560. MR2123200

Cadre, B. (2006). Kernel estimation of density level sets. *Journal of Multivariate Analysis* **97** 999–1023. MR2256570

Cuevas, A. (2009). Set estimation: Another bridge between statistics and geometry. *BEIO* **25** 71–85. MR2750781

Cuevas, A. and Fraiman, R. (1997). A plug-in approach to support estimation. *The Annals of Statistics* **25** 2300–2312. MR1604449

Cuevas, A. and Fraiman, R. (2009). Set estimation. *New Perspectives in Stochastic Geometry* **1** 374–398. MR2654684

Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Adv. in Appl. Probab.* **36** 340–354. MR2058139

Dembo, A. and Peres, Y. (1994). A topological criterion for hypothesis testing. *The Annals of Statistics* **22** 106–117. MR1272078

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York. MR1383093

Devroye, L. and Györfi, L. (2002). Distribution and density estimation. In *Principles of Nonparametric Learning; CISM Courses and Lectures No. 434* (L. Györfi, ed.) 211–270. Springer Verlag, Vienna. MR1987660

Devroye, L. and Lugosi, G. (2002). Almost sure classification of densities. *Journal of Nonparametric Statistics* **14** 675–698. MR1941709

Erdős, P. (1945). Some remarks on the measurability of certain sets. *Bull. Am. Math. Soc.* **51** 728–731. MR0013776

Gorban, A. N., Kégl, B., Wunsch, D. C. and Zinovyev, A., eds. (2007). *Principal Manifolds for Data Visualization and Dimension Reduction. Lecture Notes in Computational Science and Engineering* **58**. Springer, Berlin Heidelberg. MR2447219

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, New York. MR1666064

Lang, R. (1986). A note on the measurability of convex sets. *Archiv der Mathematik* **47** 90–92. MR0855142

Lee, J. A. and Verleysen, M. (2007). *Nonlinear Dimensionality Reduction. Information Science and Statistics*. Springer, New York. MR2373983

Li, S. (2011). Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics* **4** 66–70. MR2813331

Mason, D. M. and Polonik, W. (2009). Asymptotic normality of plug-in level set estimates. *Annals of Applied Probability* **19** 1108–1142. MR2537201

Pateiro-López, B. and Rodríguez-Casal, A. (2009). Generalizing the convex hull of a sample: The R package alphahull. *Journal of Statistical Software* **34** 1–28.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics* **22** 2031–2050. MR1329181

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling*. Springer, New York. MR1707286

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics* **23** 855–881. MR1345204

Rigollet, P. and Vert, R. (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli* **15** 1154–1178. MR2597587

Rodríguez-Casal, A. (2007). Set estimation under convexity-type assumptions. *Ann. Inst. H. Poincaré Probab. Statist.* **43** 763–774.

Royden, H. L. (1968). *Real Analysis*. Macmillan, New York. MR0151555

Schick, A. (1997). On U-statistics with random kernels. *Statistics and Probability Letters* **34** 275–283. MR1458022

Scott, C. D. and Nowak, R. D. (2006). Learning minimum volume sets. *Journal of Machine Learning Research* **7** 665–704. MR2274383

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons. MR0595165

Shao, C., Huang, H. and Wan, C. (2007). Selection of the suitable neighborhood size for the ISOMAP algorithm. In *Proceedings of International Conference on Neural Networks* 300–305.

Steinwart, I., Hush, D. and Scovel, C. (2006). A classification framework for anomaly detection. *Journal of Machine Learning Research* **6** 211. MR2249820

Tenenbaum, J. B., de Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *The Annals of Statistics* **25** 948–969. MR1447735

Vert, R. and Vert, J. P. (2006). Consistency and convergence rates of one-class SVMs and related algorithms. *Journal of Machine Learning Research* **7** 817–854. MR2274388

Willett, R. M. and Nowak, R. D. (2007). Minimax optimal level-set estimation. *IEEE Transactions on Image Processing* **16** 2965–2979. MR2472804